Next Generation Population Genomics in the Guppy (*Poecilia reticulata*)

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl. Inform. Eva-Maria Willing
aus Vreden

Tübingen
2011

Tag der mündlichen Qualifikation:	25.01.2012
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Daniel Huson

Erklärung

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind durch Angaben der Quellen kenntlich gemacht wurden. Die Doktorarbeit besteht teilweise aus Auszügen eigener Publikationen oder Publikationen, die derzeit in Vorbereitung sind.

Tübingen, November 2011
Eva-Maria Willing

Acknowledgements

First of all I would like to thank Prof. Dr. Christine Dreyer for her supervision, trust and support the last four years. I have developed a lot of new skills during this time and I appreciate the freedom I had in selecting the topics my research. I would also like to thank Prof. Dr. Detlef Weigel for his support, discussions and the opportunity to do my PhD in his department at the MPI. It was a very inspiring environment providing endless opportunities. Furthermore, I would like to thank Prof. Dr. Daniel Huson, who took the supervision at the university and also helped me with discussions and suggestions during my thesis

I would like to thank all members of the guppy group for providing such a nice working environment. Especially, Margarete Hoffmann for her excellent wet lab work, support and discussions. Without her I would have not been able to perform my research, because I would not have had any data to analyze. Namita Tripathi, who worked in the group the first two years during my thesis and who was a very helpful and inspiring colleague during that time. Verena Kottler, who started as a new colleague, but has become a very close friend during this time sitting always at the left site of me only 50 cm away. I have to thank her for all the supporting words, cheer ups and coffees during the bad days. Andrea Sprecher for her help at first as a HIWI and then later as a diploma student. In addition, I also would like to say thank you to all members of the Weigel Lab for a very productive and friendly environment and of course my collaborators who shared their samples and knowledge with me.

Furthermore, I would like to say thanks to my friends at the MPI. Especially, Chris, Dave and Sebi, who went for lunch with me almost every single day during this time. It was always a very relaxing, entertaining and informative hour. Steffi for all the nice coffee breaks we had together.

I would also like to say thank you to all my friends in Tübingen. Especially, I would like to say thank you to Marc for his calm support all these years and Vera, who has been a very supportive friend since my first day in Tübingen.

Last but not least, I say thank you to my parents, Alfons and Mathilde, and my siblings Mechthild, Rudolf and Andrea for their kind support during my whole life.

In accordance with the standard scientific protocol, I will use the personal pronoun “we” to indicate the reader and the writer, or my scientific collaborators and myself.

Table of Contents

Zusammenfassung

Populationsgenetische Studien beschreiben die Verteilung von Allelfrequenzen mit dem Ziel, deren Veränderung über die Zeit abzuleiten, woraus wiederum auf den demographischen Werdegang natürlicher Populationen rückgeschlossen werden kann. Darüber hinaus versuchen sie das Phänomen der Adaptation und der Artenbildung zu erklären. Bis vor kurzem basierten Studien in Nicht-Referenzorganismen auf nur wenigen genotypisierten Loci, da die Neuentwicklung einer großen Anzahl von Markern sehr kostspielig war und darum außerhalb der Möglichkeiten der meisten Forschungsprojekte lag. Bisher waren nur in etablierten Modelorganismen mit bekanntem Referenzgenom „populationsgenomische“ Studien, die das genomweite Muster von Sequenzvariationen innerhalb und zwischen nahverwandten Populationen und Spezies untersuchen, möglich. Die Verfügbarkeit von neuen Sequenziertechnologien der nächsten Generation (NGS) hat nicht nur grosse Fortschritte in der Genomforschung ermöglicht, sondern auch die Entwicklung genomweiter Marker erleichtert. Der Guppy (*Poecilia reticulata*) ist ein wichtiger Modelorganismus in der ökologischen Genetik. Die Anpassung von Guppies in Hinblick auf Verhalten, Morphologie und Lebensweg in gegensätzlichen oberen und unteren Flussläufen wurde ausführlich untersucht. Guppies sind in der Lage sich schnell an eine neue Umgebung anzupassen, was vermutlich an der hohen natürlichen Variation liegt. Bisher war es unmöglich genomweite Analysen genetischer Variabilität durchzuführen oder nach Regionen im Genom, die einen Selektionsvorteil aufweisen, zu suchen. Diese Arbeit beschreibt den Übergang von Populationsgenetik zu Poplationgenomik im Guppy. Zuerst untersuchen wir die Populationsstruktur in einer Auswahl natürlicher Populationen und suchen zum ersten Mal nach Regionen mit Selektionsvorteil mittels eines genomweiten Satzes von genetisch kartierten Single Nukleotid Polymorphismus (SNP) Markern. Durch die Simulation von Populationen konnte ich abschätzen welchen Einfluss die Stichprobengröße einer Population und die Anzahl der Marker auf unterschiedliche Schätzer von genetischer Differenzierung haben. Ich demonstriere wie NGS genutzt werden kann, um die Gene in Regionen von Interesse, auch ohne Referenzgenom, zu identifizieren und am Ende zeige ich wie Sequenzierung von DNA Abschnitten neben Restriktionsschnittstellen (RAD-seq) die genomweite Entwicklung von SNP Markern im Guppy ermöglicht.

Abstract

Population genetic studies estimate allele frequency distributions and the change of these frequencies over time in order to infer the demographic history of natural populations. Such studies aim to explain how adaptation and speciation have occurred. Until recently, inferences in non-reference taxa have been based on very few loci due to the high cost of developing a large set of markers *de-novo*. Only in established model organisms with a known reference genome was it possible to study genome-wide patterns of sequence variation. However, the advent of Next Generation Sequencing (NGS) technologies has revolutionized the field of whole genome research, and facilitated the development of genome-wide genetic markers.

The guppy (*Poecilia reticulata*) is an important model organism in ecological genetics. Adaptation of guppies to contrasting upland and lowland habitats has been extensively studied with respect to behavior, morphology and life history. Guppy populations are able to adapt rapidly to new environments, presumably due to their high level of standing natural variation. However, it was previously not possible to deduce a genome-wide picture of genetic variability and to scan for the causative genomic regions under selection. In this thesis, I will describe our efforts to move from population genetics to population genomics in the guppy. This was achieved by first using a genome-wide set of genetically mapped single nucleotide polymorphism (SNP) markers for the analysis of population history and then, for the first time, to check for regions under selection in the guppy genome. By simulating populations, I assessed the effects of sample size and marker number on the various estimates of genetic differentiation. I will show how NGS can be used to identify genes in genomic regions of interest without an available reference genome and, finally, I will describe how restriction associated DNA sequencing (RAD-seq) facilitates the development of genome wide SNP markers in the guppy.

1. Introduction

1.1 The field of population genetics

Population genetics is a field that studies allele frequency distribution and their change over time in natural populations (Weir, 1996). Differences in allele frequency distributions among natural populations from the same species give insights into the demographic history and structure of these populations. Measuring allele frequencies at many loci within a population can be considered a description of the population. Changes of allele frequencies within natural populations over time are at the core of explaining evolution and speciation.

There are four main evolutionary processes causing changes in allele frequencies: genetic drift, gene flow, natural selection and *de novo* mutation (Hartl and Clark, 1997). Genetic drift describes the process of random sampling of parental alleles that occurs from one generation to the other. Genetic drift can lead to genetic differentiation of separate populations within the same species. Gene flow describes the exchange of alleles among populations, most likely by migration of fertile individuals from one population to another. Gene flow between populations reduces the genetic differentiation. Mutations are the origin of all genetic variation. They generate new alleles within a population by altering the DNA sequence, increasing the genetic variability. Most of the mutations have no effect on the fitness of an individual and are therefore called neutral. However, some mutations do have an effect on the fitness of an individual within a population. This effect can either result in an enhancement of fitness (advantageous mutations) or in a reduction of fitness (deleterious mutations). If a mutation has an effect on the fitness, natural selection can act on it. In the case of advantageous mutations it causes the allele to become more common within a population. In the case of deleterious mutations natural selection will cause the allele to become less common within a population. Whereas genetic drift and gene flow are random processes affecting the whole genome in the same way, natural selection is a directed process, affecting particular loci in the genome. However, if environmental conditions change over time, previously neutral mutations can become e.g. advantageous and the allele carrying the mutation that was rare before becomes more common within the population due to positive selection. A textbook example is the peppered moth (*Biston betularia*) (van't Hof, et al., 2011). Within this species there are two alternative alleles for coloration, one causing light coloration and one causing dark coloration. Originally, the allele for light coloration was the common one within the population. The light coloration effectively camouflaged the moths against the light-colored trees they rested on protecting them from predation. However, due to widespread pollution during the Industrial Revolution in England the trees, which peppered moths rested on, became blackened by soot. Now, the allele for dark coloration is advantageous, because it provides the ability to hide on the darkened trees. However, without the genetic variability at this locus, the species would not have been able to adapt to the environmental change. A population of individuals can only adapt, if it contains genetic variability, which is the core of population genetic studies.

1.2 From population genetics to next generation population genomics

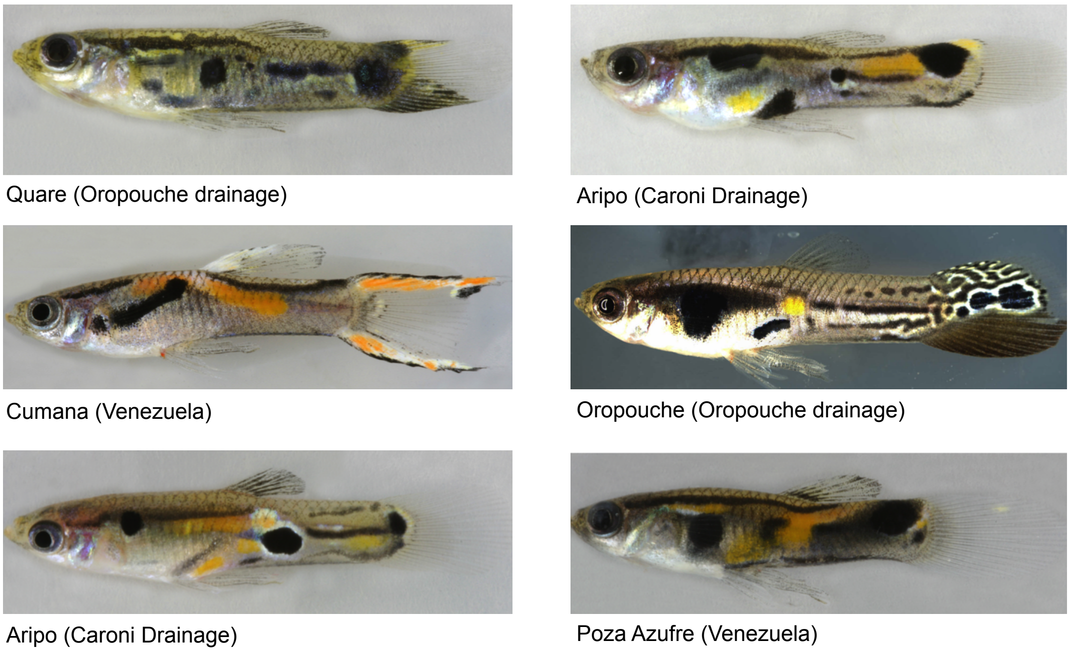
Until recently, most studies on wild populations of non-reference species used moderately large numbers of samples per population (> 20), but only a small number of genetic markers (< 20). The most widely used markers are amplified fragment length polymorphisms (ALFPs), microsatellites and single nucleotide polymorphisms (SNPs) (Luikart, et al., 2003). However, the term "population genomics" was already used more than ten years ago in a publication about human disease genetics by Gulcher and Sefansson (1998). Luikart et al. (2003) defined population genomics "as the simultaneous study of numerous loci or genome regions to better understand the roles of evolutionary processes (such as mutation, random genetic drift, gene flow and natural selection) that influence variation across genomes and populations". They suggested genotyping tens to hundreds of genomic markers in order to do a population genomic study. Yet, molecular resources in ecological model organisms sufficient for genome wide marker development have been rare at that time and population genomic studies have only been feasible in a number of reference organisms, like humans (Salmela, et al., 2008). The development of a large number of genetic markers for new species typically involved marker discovery for which sufficient molecular resources were needed, assay development for each marker, and proof of the assays in a screening population before full deployment across large populations. This process was costly (in time and research funding) and usually resulted in generation of very few (tens) of working markers. Still, genome-wide data sets have the potential to improve the inference of population parameters and to reliably reconstruct population demography and evolutionary history. Moreover, they can provide a better understanding of adaptive evolution (Luikart, et al., 2003; Narum, et al., 2008). Genome-wide SNP marker sets have already proven useful in humans revealing detailed genome-wide perspectives on phylogeographic relationships (Salmela, et al., 2008 and citations there in) and evidence for genes under natural selection (Akey, et al., 2002). However, the examination of SNPs in wild populations addressing evolutionary, ecological or conservation issues is still limited to a few examples including wolves (Seddon, et al., 2005), lizards (Rosenblum and Novembre, 2007) and salmon (Narum, et al., 2008).

Single nucleotide polymorphisms (SNPs) represent the most abundant source of variation in the genome of most organisms. Their distribution throughout the entire genome at high density, well-established models for handling mutation rates and error rates, and the methods for high throughput genotyping make them appealing for population genetic studies at the whole genome level (Narum, et al., 2008). Yet, due to the lack of genomic information for non-reference organisms, the development of large SNP marker sets is rather expensive in such species. Therefore, microsatellites and AFLPs have been the markers of choice for molecular studies in ecology and evolution. Microsatellites are appealing because of their high variability and consequently high information content, which can be four to ten fold higher for some multi-allelic microsatellites as compared to bi-allelic SNPs (Morin, et al., 2004). However, analyses of microsatellite data suffer from complicated mutation models, high incidence of homoplasy, potentially error-prone assays and low genotyping throughput, making a genome-wide analysis with microsatellites very difficult (Di Rienzo, et al., 1994; Estoup, et al., 1995; Hedrick, 1999; Hoffmann and Amos, 2005; Miller, et al., 2002). AFLPs are cheap to develop and allow the analysis of thousands of polymorphisms spread across the genome without having access to DNA sequence information. However, the relatively poor per-locus type of genetic information is a major disadvantage (Bensch and Akesson, 2005). In most studies it is impossible to separate between dominant homozygous (1/1) and heterozygous (1/0).

In recent years, several novel high-throughput sequencing platforms have entered the market. The most commonly used are the SOLiD system by Applied Biosystems (www.appliedbiosystems.com), the Solexa technology, now owned by Illumina (www.illumina.com) and the 454 platform (Margulies, et al., 2005), now owned by Roche (www.roche.com) (for a description of all three technologies see (Mardis, 2008). These Next Generation Sequencing (NGS) technologies have revolutionized the field of genome research, at first by allowing cheap re-sequencing projects for organisms with an already existing reference genome. But recently more and more methods have been developed incorporating NGS to analyse also non-reference organisms, taking advantage of improvements such as longer read lengths and paired-end (PE) reads. Therefore, it is now much more difficult to understand the ecology of established genetic model organisms than to generate new resources required for the genetic analyses of established ecological model organisms (Tautz, et al., 2010). Given the availability of ultra-high-throughput sequencing, one strategy for marker development might be to just sequence completely the genomes of several of the target organisms, and identify SNPs by comparing these total data sets. However, this is still not feasible for eukaryotes with genomes that are hundreds or thousands of megabase pairs in size. Furthermore, *de-novo* assemblies of large genomes from very short reads remain difficult, in spite of recent improvements in assembly algorithms (Gnerre, et al., 2011). However, more and more methods are becoming available allowing *de-novo* discovery and genotyping of genome-wide SNP markers in many individuals. The probably most promising development is restriction site associated DNA (RAD) sequencing, which allows large amounts of sequence data to be generated for mapping or population genetic studies, but without the need for prior identification of SNP sites (Baird, et al., 2008). Further, techniques for target enrichment are now available for NGS in order to sequence particular genome regions of interest (Summerer, 2009).

1.3 The Guppy as model organisms in ecological genetics

Guppies (*Poecilia reticulata*) are native to freshwater habitats in the North Eastern coastal range of South America, including the offshore island of Trinidad (Magurran, 2005). They are long-standing models for studies on ecological genetics and most fieldwork has been performed in the Northern Mountain Range of Trinidad. On the Southern slope of the mountain range are the Oropouche and Caroni drainages, which have been separated by a watershed divide for 600,000 to 1,000,000 years (Carvalho, et al., 1991; Fajen and Breden, 1992). Before separation from the South American mainland, the present Caroni drainage (including the portions along the northern coast draining directly into the sea) and Oropouche drainage were most likely connected to North and South flowing arms of the Orinoco, respectively, giving rise to at least two distinct arcs of colonization that are still manifested in a distinct freshwater fauna. This proposal is known as the "two arcs hypothesis" (Kenny, 1988; Magurran, 2005). However, it has been noted that there are substantial differences between the North flowing rivers and both the Oropouche and Caroni. While mullets, gobies and freshwater prawns prevail as main guppy predators in the Antillean fauna of the Northern drainages, guppies are preyed upon by the typical South American mainland cichlids and characins in the Caroni and Oropouche (Magurran, 2005; Reznick and Bryga, 1996) and citations therein). Studies on ecological genetics revealed an amazing degree of phenotypic natural variation in male nuptial ornaments (Figure 1.1), courtship behavior and life history traits within and among wild guppy populations. This phenotypic variation is shaped by both natural and sexual selection and has been primarily studied in the context of adaptations to habitats that differ in the presence of predators (Endler, 1995; Houde, 1997; Magurran, 1998). Usually, waterfalls separate upper river ranges from lower river ranges and prevent the migration of big predators upstream (Figure 1.2).



Quare (Oropouche drainage)

Aripo (Caroni Drainage)

Cumana (Venezuela)

Oropouche (Oropouche drainage)

Aripo (Caroni Drainage)

Poza Azufre (Venezuela)

Figure 1.1: Phenotypic natural variation in male nuptial ornaments found among individuals in different drainages.

Consequently, headwaters are most of the time low predation habitats, whereas downstream sites are usually high predation habitats. High and low predation populations differ in a broad suite of phenotypic traits that appear to be adaptively significant, namely the conspicuousness in male coloration (Endler, 1978; Endler, 1980; Endler, 1983; Godin and McDonough, 2003; Houde, 1997), behavior (Breden, et al., 1987; Ghalambor, et al., 2004; Kelley and Magurran, 2003; Magurran and Seghers, 1990; Magurran, et al., 1992; Magurran, et al., 1995; O'Steen, et al., 2002; Seghers, 1974), life history traits (Reznick and Bryga, 1987; Reznick, et al., 1996b) and parasite resistance and load (Martin and Johnson, 2007 ; van Oosterhout, et al., 2003; van Oosterhout, et al., 2007). Where barrier waterfalls separate upper and lower stream populations, contemporary gene flow between upper and lower habitats occurs primarily in the downstream direction (Barson, et al., 2009; Crispo, et al., 2006; Shaw, et al., 1991). While different upper river ranges within drainages remain isolated from each other, there is gene flow among different lower river habitats (Barson, et al., 2009). Artificial introductions are a powerful tool for experimentally determining the selective forces driving adaptive divergence and studying adaptation in real time. Most introduction experiments have been performed within the same

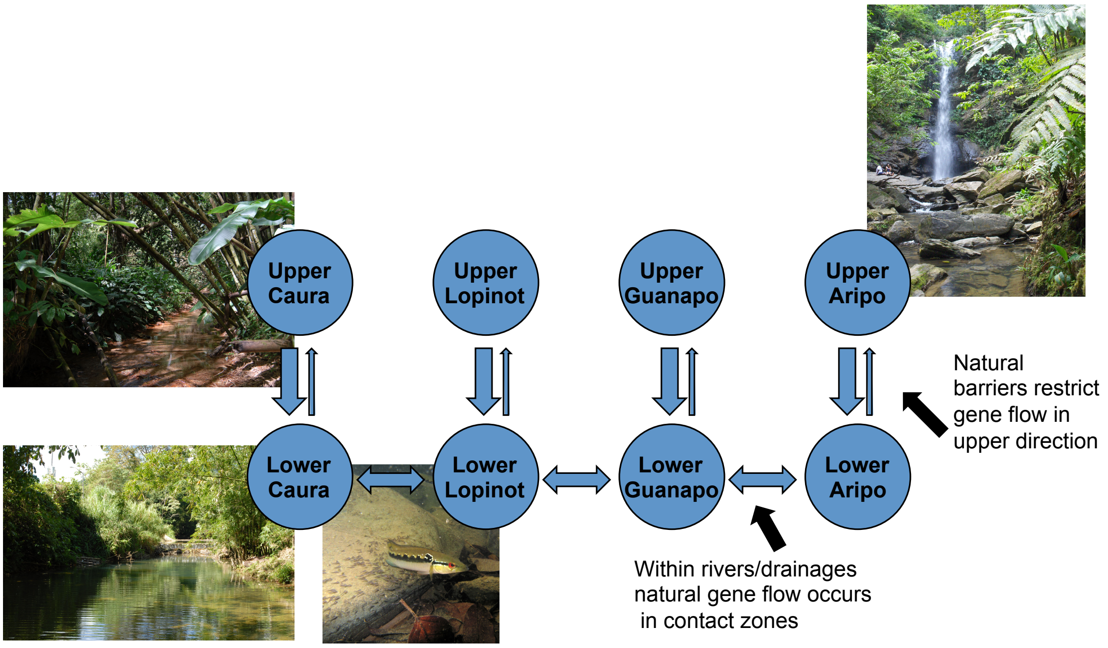

Upper Caura
Upper Lopinot
Upper Guanapo
Upper Aripo
Natural barriers restrict gene flow in upper direction
Lower Caura
Lower Lopinot
Lower Guanapo
Lower Aripo
Within rivers/drainages natural gene flow occurs in contact zones

Figure 1.2: Migration model among rivers within the same drainage. Different upper and lower rivers are depicted by blue circles (river names are taken from the Caroni drainage as an example). Arrows show the gene flow occurring between different river parts and the arrow size reflects the amount of gene flow. Photos taken by Paul Bentzen and Eva-Maria Willing

river basin to sites that were previously devoid of guppies. John Endler transferred guppies within the Aripo in 1976 (Endler, 1980), David Reznick within the El Cedro in 1981 (Reznick and Bryga, 1987) and also in 1996 from the Lower Yarra to the adjacent Damier River (Karim, et al., 2007); reviewed by (Magurran, 2005). In contrast, about 50 years ago Haskins transferred about 200 guppies from the Caroni drainage to the Oropouche drainage, namely from lower Guanapo to upper Turure (Magurran, 2005; Shaw, et al., 1992). Striking changes in male color patterns and life history traits were reported shortly after such transfers (Endler, 1980; Karim, et al., 2007; Reznick, et al., 1990) and the rapid evolutionary changes observed suggest that some natural guppy populations must harbor substantial standing variation to realize this evolutionary potential within a limited number of generations.

Previously, limited numbers of mitochondrial markers (Alexander and Breden, 2004; Alexander, et al., 2006; Fajen and Breden, 1992), allozymes (Carvalho, et al., 1991; Shaw, et al., 1991) and microsatellites (Crispo, et al., 2006; Suk and Neff, 2009) have been used to study population genetics and phylogeographic history in guppies. These analyses have indicated significant genetic divergence among guppies from the different drainages and also considerable substructure among populations within drainages. However, inferences about the phylogeographic relationships among the studied populations are not always congruent.

Studies using allozymes and mtDNA have revealed marked genetic divergence between populations from the Oropouche drainage and the Caroni and Northern drainages, supporting the hypothesis of two major lineages of guppies in Northern Trinidad (Alexander, et al., 2006; Carvalho, et al., 1991; Fajen and Breden, 1992). However, a recent study based on seven microsatellites found that populations from the Northern drainages are all highly differentiated from populations in either the Caroni or Oropouche drainage (Suk and Neff, 2009). The same study also suggested that guppies from the Aripo River within the Caroni drainage have a genetic signature that is more similar to populations from the Oropouche drainage, in contrast to previous results indicating that populations from the Aripo river are more closely related to other populations within the Caroni drainage (Alexander, et al., 2006; Carvalho, et al., 1991; Fajen and Breden, 1992).

Therefore, a genome-wide picture of standing genetic variation might refine the reconstruction of the evolutionary history of populations and would enhance the understanding of adaptive evolution in different guppy populations.

1.4 Outline of the thesis

At the beginning of this thesis, the only molecular resources available for the guppy were a library of expressed sequence tags (ESTs) (Dreyer, et al., 2007) and randomly sequenced bacterial artificial Chromosome (BAC) end sequences. These resources were used to develop approx. 1,000 single nucleotide polymorphism (SNP) markers in order to generate a genetic map and to perform QTL analyses (Tripathi, et al., 2009).

Since a genome-wide picture of standing genetic variation has not been conducted so far, these approx. 1,000 nuclear markers were genotyped in 239 individuals from 37 sites in Trinidad and Venezuela. Chapter 2 is based on Willing et al. (2010) and describes how used this dataset was explored to make inferences about demographic history of these populations using modern population genetic methods and custom methods specially implemented for this analysis. In addition, the genome-wide set of genetically mapped SNP markers was used to scan for regions under selection in the guppy genome using F_{ST} outlier methods.

In Chapter 2, only a very small number of individuals sampled per population were used (on average 5.7). It is commonly thought that large sample sizes are required in order to reliably infer F_{ST} and that small sample sizes lead to overestimation of genetic differentiation. Chapter 3 is based on Willing et al. (*in preparation*) and examines whether a large number of genetic markers can substitute for small sample sizes when estimating F_{ST}. The behavior of three different estimators that infer F_{ST} and that are commonly used in population genetic studies was tested.

However, 1,000 SNP markers distributed equally over the guppy genome with a size of approx. one gigabase leaded to a SNP density of approx. one SNP per megabase. This density is far too low in order to detect the regions under selection due to predation pressure. It is likely, that the allele variants, that are under selection, already existed in the last common ancestor of the Trinidadian populations. Consequently, linkage blocks are expected to be small and a high marker density is required to get the resolution for detecting signals of selection. Therefore, we were looking for a cost effective method to develop more genomic markers. Sequencing restriction site associated DNA (RAD) with Illumina was developed in stickleback for rapid detection of new SNP markers.

In Chapter 4, a new approach is demonstrated that uses paired-end RAD-seq strategy to produce extended contigs flanking a restriction site and how these can be used to develop thousands of new SNP markers. This Chapter is based on Willing et al. (2011).

As mentioned before, the genome-wide set of SNP markers was used to detect genomic regions under selection. However, the scored SNPs alone do not give any hints about the genes linked to these SNPs. Since there is no reference genome sequence for the guppy available (which is true for most taxa in ecological genetics), another strategy had to be designed in order to investigate genomic regions of interest (e.g. under selection). Bacterial artificial chromosomes (BACs) contain the (possibly large) genome of an organism in comparatively small pieces and BAC libraries can be screened for genomic regions of interest. Chapter 5 examines the feasibility of using the Illumina technology to sequence and assemble a library made of pooled BAC DNA.

2. Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies

Previously, genotyping assays were developed for over 1,000 SNPs that were used to construct a complete linkage map of the guppy (Dreyer, et al., 2007; Tripathi, et al., 2009). The study described in this Chapter was published in Molecular Ecology in 2010 (Willing et al, 2010) and surveys the value of these genome-wide markers in the analysis of genetic variation within and among naturally occurring guppy populations. The study was conceived by Christine Dreyer and Detlef Weigel. Paul Bentzen, Cock van Oosterhout, Joanne Cable and Felix Breden provided the population samples analyzed. Margarete Hoffmann prepared the DNA needed for the genotyping. I conceived and performed all the analyses described. Christine Dreyer, Paul Bentzen, Joanne Cable, Cock van Oosterhout, Felix Breden and Detlef Weigel helped with the interpretation of the results (see also Contributions).

The genotypes of 239 individuals representing 37 sample sites from a wide geographic range covering Trinidad and Venezuela were ascertained. I subjected the resulting dataset to modern population structure analysis by three different individual-based methods that do not require *a priori* knowledge of predefined populations. I completed the analyses by the estimation of well-established population statistics including F_{ST}, AMOVA and expected heterozygosity (H_e). By comparing my findings to previous work, which employed other marker types, I assessed the performance of our SNP data set and tested whether our SNPs are informative in a large number of different populations. A genome wide picture of standing variation may extend and refine previous results, but it can also clarify previous incongruent conclusions about phylogeographic relationships and provide new knowledge on phylogeographic history and undetected admixture events leading to some new interpretation. Additionally, the data gave first indication of potential regions under selection by scanning the genome with $\mathrm{F_{ST}}$ outlier methods (Lewontin and Krakauer, 1973). Under selective neutrality, genetic drift and gene flow determine the allele frequency divergence of all loci across the genome and therefore determine F_{ST}. Since this random process is expected to affect all loci in a similar way, SNPs with extremely high or low F_{ST} values may be strong candidates linked to selectively important loci. We hypothesized that neutral forces, such as random genetic drift, are mainly responsible for the large genetic divergence among guppies from different drainages. However, the freshwater habitats of the Northern, Caroni and Oropouche drainages exhibit many differences, including different types of predators that might drive adaptation (Magurran, 2005). Therefore, I scanned this novel data set in order to find evidence for genomic regions under adaptive natural selection in guppies originating from the Oropouche, Caroni and Northern drainages.

2.1 Material and Methods

2.1.1 Sample collection and SNP analysis

The 239 individuals were sampled from five major geographical regions in Trinidad and Venezuela, namely the Caroni drainage (North-West Trinidad), Oropouche drainage (North-East Trinidad), the Northern drainages of Trinidad (Yarra, Marianne and Paria), South-West Trinidad and Eastern Venezuela (Figure 2.1, Table 2.1). The samples can be hierarchically classified. The top level represents the five major geographical regions that are presently separated by watershed divides or by the sea. The next level corresponds to samples originating from different river basins in the separate geographical regions, and the lowest levels are the different sampling sites within the same river, e.g. upstream and downstream sites (Table 2.1). In total, samples from 37 different sites were available, with N=2 to 14 (mean 5.7, see Discussion) individuals per sample site. Although the sample sizes per population were small, the total sample sizes of populations within four of the five major geographic regions were moderately large (Oropouche drainage N=42, Caroni drainage N=70, Northern drainages N=63 and Venezuela N=20). In order to minimize the chance of sampling closely related individuals, samples from the same spot at one sampling site were avoided. Of the samples 187 came directly from the wild, and 95 of these were parts of larger samples from 16 sites, which had previously been genotyped for nine microsatellites to confirm that these populations were in Hardy-Weinberg equilibrium. The analysis included 52 progeny from wild caught guppies that were kept for up to 10 - 15 generations (3 to 4 years) in the laboratory (see Table 2.1). Lower Oropouche, upper Quare and Cumaná samples (labeled with an asterisk in Table 2.1) had each been separated from natural samples and subdivided into families reflecting different color patterns before breeding in small community tanks.

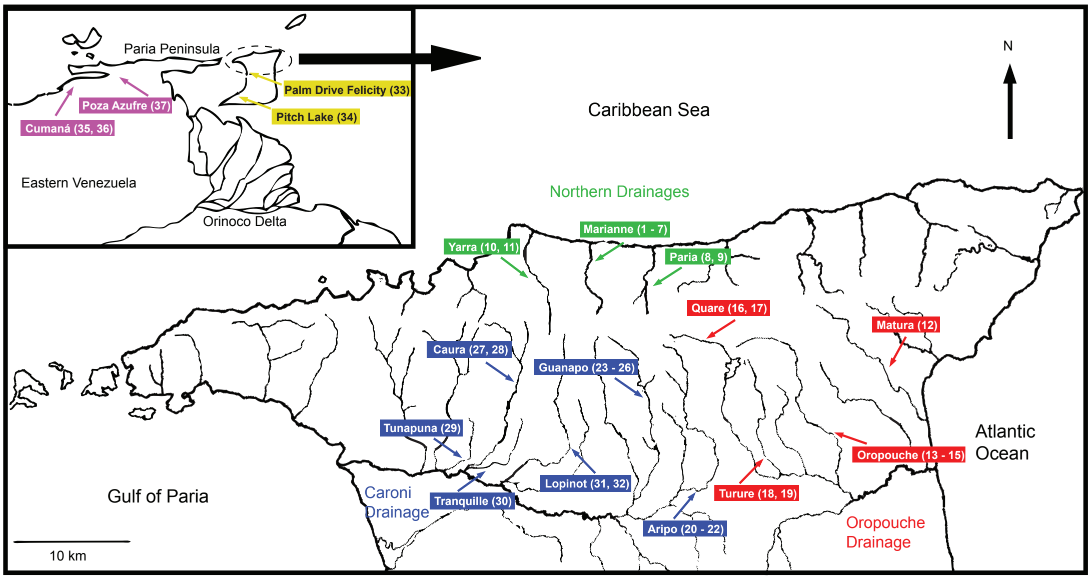


Figure 2.1: Map of East Venezuela and Trinidad showing sample sites. Map in inset shows locations in East Venezuela and South-West Trinidad. Sites from which the fish were collected are labeled and numbered as in Table 2.1. Color codes: Venezuela – purple, South-West Trinidad – yellow, Northern drainages – green, Caroni drainage – blue, Oropouche drainage – red.

To provide the best estimate of the natural variation originally contained in the sample I retained one to four specimens from each family for the final analysis.

Genomic DNA was isolated from the tail muscle of each fish using a Qiagen DNeasy96 kit (Catalogue No.69582) according to the manufacturer's instructions. DNA was adjusted to a concentration of 20 ng/µl and 2.5 ng per SNP assay was provided.

2.1.2 SNP genotyping and evaluation

Genotyping assays for 1005 polymorphic markers were designed and performed using Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF) (by Sequenom (San Diego, CA, USA), for details see (Vignal, et al., 2002)), based on sequence information gained from genomic re-sequencing of two populations, namely upper Quare and Cumaná (Tripathi, et al., 2009).

After exclusion of markers whose assays failed and 104 markers that were not polymorphic (< 1% frequency of the minor allele) a set of 866 SNPs remained for further analysis. This marker set had a mean missing data rate of 10.3% across SNPs, 8.8% across individuals and 8.9% across populations. I used these 866 SNPs as input for the Neighbor-Net analysis. To minimize the bias in the inference of admixture coefficients by STRUCTURE, I reduced the mean missing rate per individual to below

Table 2.1: List of sampling sites.

Geographical region	River	Sample Site	Number of Individuals	Population Number	Predation Regime
Northern drainages (Northern slope of North Trinidad)	Marianne	M1	6	1	Low
		M11	6	2	Low
		M15	5	3	High
		M2	6	4	Low
		M8	6	5	Low
		Mid Marianne 1	4	6	High
		Mid Marianne 2	4	7	High
	Paria	Paria	6	8	Low
		Paria 7	6	9	Low
	Yarra	Lower Yarra	14	10	High
		Upper Yarra	14	11	Low
Oropouche drainage (Eastern part of Southern slope of North Trinidad)	Matura	Lower Matura	4	12	High
	Oropouche	Lower Oropouche	4	13	High
		Lower Oropouche*[1]	12	14	High
		Upper Oropouche	4	15	Low
	Quare	Lower Quare	6	16	High
		Upper Quare*[2]	12	17	Low
	Turure	Lower Turure	12	18	High
		Upper Turure	10	19	Low
Caroni drainage (Western part of Southern slope of North Trinidad)	Aripo	Lower Aripo	14	20	High
		Rapsey Pool	4	21	Low
		Upper Aripo	6	22	Low
	Guanapo	Lower El Cedro	6	23	High
		Upper El Cedro	6	24	Low
		Lower Guanapo	4	25	High
		Upper Guanapo	4	26	Low
	Caura	Lower Caura	4	27	High
		Upper Caura	4	28	Low
	Tunapuna*	Tunapuna*	2	29	Low
	Tranquille*	Tranquille*	8	30	High
	Lopinot	Lower Lopinot	4	31	High
		Upper Lopinot	4	32	Low
SW Trinidad	Palm Drive Felicity		4	33	High
	Pitch Lake		4	34	Low
Venezuela (Mainland)	Cumaná	Armando Pou	2	35	High
		Central Cumaná*[3]	13	36	High
	Poza Azufre	PV6*	5	37	High

NOTE: *Individuals that have been separated from natural populations and subdivided into families (see Material and Methods). For the analysis individuals from different families have been pooled to provide the best estimate of variation originally contained in the sample yielding [1] four Oro209, two Oro201, two Oro4-2 and four Oro2 individuals [2] four Quare6, four Qua6_II-203, two Qua6_II-206, one Qua6_II-215-3 and one Qua6_3-2 individuals, [3] three CCBlue, four CCELB and six CCFR individuals.

10% (Pritchard and Wen, 2002) by excluding 60 markers with a missing call rate of > 40%. With the remaining 806 markers the mean missing rate across individuals was 4.8% and only 11 individuals had a missing call rate between 10% and 17%. The mean missing rate across SNPs and across populations was 5.9% and 4.8%, respectively. Since the input format for the software EIGENSOFT demands positional information of all markers used, for PCA I could only use 720 markers with information about their position on the genetic map (Tripathi *et al.* 2009).

2.1.3 Neighbor-Net

Since relationships among populations may not conform to a tree-like pattern due to potential gene flow (Nordborg, et al., 2005) and shared ancestral polymorphisms, I performed a phylogeographic analysis using the method Neighbor-Net (Bryant and Moulton, 2004) implemented in SPLITSTREE4 (Huson and Bryant, 2006) after compiling an artificial nucleotide sequence comprising all SNPs with heterozygous SNPs coded according to IUPAC. Because all populations analyzed were considered members of the same species, I assumed that more than one mutation at a specific position had rarely occurred and used *Uncorrected_P* distance as metric and ambiguous states were handled as average matches. The Normalize option accounted for unequal distribution of missing data across individuals.

2.1.4 Bayesian analysis by STRUCTURE

For direct assessment of admixture I used STRUCTURE version 2.3 (Pritchard, et al., 2000). This Bayesian clustering approach avoids *a priori* population classifications, and instead estimates the shared population ancestry of individuals based solely on their genotypes assuming Hardy-Weinberg equilibrium and linkage equilibrium in ancestral populations. The number of predefined clusters (k) has to be optimized by the user. I ran STRUCTURE with a Markov Chain Monte Carlo (MCMC) burn in of 100,000 steps, followed by an MCMC of 100,000 steps for clustering inference. The allele frequency distribution, which is assumed to be a Dirichlet distribution, is parameterized by λ. I estimated λ in three different runs for each dataset prior to the main clustering analysis by setting $k = 2$ and fixed λ at the mean of the inferred values. I used the admixture model with correlated allele frequencies (Falush, et al., 2003) and applied the infer α option with the same α for all populations, where α

parameterizes the Dirichlet distribution used to infer admixture coefficients. Each analysis was repeated 10 times for each *k.* To find the optimal *k,* I used the Wilcoxon rank-sum test to determine whether or not likelihoods improved significantly for *k* compared to *k*-1 clusters. I calculated the symmetric similarity coefficient (*SCC*) with CLUMPP (Jakobsson and Rosenberg, 2007) between all pairs of runs for the same *k*. When multiple runs at the same *k* value produced discrepant results, I relied on the majority rule and the best mean likelihood to select the optimal result.

2.1.5 Principal Component Analysis

Principal Component Analysis (PCA) has similar power to detect population structure as STRUCTURE (Patterson, et al., 2006). In addition, an estimation of the maximal number of subpopulations that can be found within a dataset is achieved by determining the number of statistically significant principal components (PCs). I used the software package EIGENSOFT version 1.01 to perform the PCA (Patterson, et al., 2006) and to test the significance of the resulting eigenvalues and corresponding eigenvectors. I used a significance cutoff of 1% to determine the number of eigenvectors representing significant population substructure. To assign the individuals to their respective population (Paschou, et al., 2007) I applied the *k*-means clustering algorithm (Hartigan and Wong, 1979) implemented in GNU R (Team, 2008) on low-dimensional data. I performed for each *k* ten independent clustering runs with maximal 10,000,000 iterations and 10,000,000 random sets to confirm the reproducibility of the results.

2.1.6 Expected Heterozygosity, F_{ST} and Analysis of Molecular Variance

For subsequent studies of variation parameters, heterozygosity and pairwise F_{ST}, I used predefined populations named by their sample sites (Table 2.1). I excluded fish from the lower Yarra, Turure, Palm Drive Felicity and Pitch Lake since the methods used so far revealed that they deviated from the hierarchical geographical pattern. Turure samples could, however, be included as part of the Guanapo cluster without substantial changes (data not shown). In contrast, including the lower Yarra samples as part of the Northern drainages decreased apparent genetic diversity between the Caroni and Northern drainages and increased genetic diversity between samples from the Yarra River and the remaining Marianne and Paria River. Considering the samples from Palm Drive Felicity and Pitch Lake as a fifth geographical region representing South-West Trinidad, did not change the overall findings, but the small sample size from this region led to a lack of power.

I performed an analysis of molecular variance (AMOVA) at different hierarchical levels using ARLEQUIN version 2.001 (Excoffier, et al., 2005). I estimated expected heterozygosity (H_e) as described (Excoffier, 2007) assuming Hardy-Weinberg equilibrium. I implemented a script in GNU R for this task. Confidence intervals were determined by 1,000 bootstraps over loci. I used the Wilcoxon rank-sum test to determine whether or not heterozygosity levels differ significantly between upper and lower river habitats, using GNU R (Team, 2008). Pairwise F_{ST}-values between subpopulations were estimated using a formula that has asymptotically minimal variance and was recently suggested by Reich *et al.* (2009, see Appendix for details). I implemented this formula in Java in order to infer pairwise F_{ST}-values (see Chapter 3). I performed a permutation test with 10,000 permutations in order to test for significance.

2.1.7 Scanning for Selection

To identify loci subject to selection I used two different programs based on different outlier approaches. First I used the program FDIST2 (Beaumont and Balding, 2004; Beaumont and Nichols, 1996) that is based on a summary statistic approach. It calculates F_{ST} for each sampled locus and then uses coalescent simulations to generate a null distribution of F_{ST} values based on an infinite island model for populations and an infinite allele model for polymorphisms (Beaumont and Nichols, 1996). We simulated the neutral distribution of F_{ST} with 50,000 iterations. The second program BAYESFST (Beaumont and Balding, 2004) relies on a Bayesian regression model implemented via a Markov Chain Monte Carlo (MCMC). It estimates three different parameters describing the locus effect (α_i), the population effect (β_j) and the interaction between both effects (γ_{ij}). I focused on the posterior distribution of the locus-effect parameters where a positive α_i suggests that locus *i* is subject to adaptive selection, whereas a negative α_i suggests balancing selection. The probability densities for F_{ST} values were obtained with the assumption of independent, lognormal (1, 1.8, 0.5) prior distributions for the α_i, β_i, and γ_{ij} (Beaumont and Balding, 2004).

To adjust the false positive rates of the two methods, I adapted the confidence levels by 10-fold to 90% for BAYESFST and 99% for FDIST2, as previously suggested by Beaumont & Balding (2004). In order to control for type I errors due to multiple testing, I repeated the simulations ten times with each program and reported only those loci detected in all ten runs at the appropriate significance level. Note that BAYESFST itself deals with the problem of multiple testing through the prior distribution of the regression parameter for the locus effect (α_i) (Beaumont and Balding, 2004).

2.2 Results

2.2.1 Clustering Analysis and Admixture Patterns

For a first analysis of guppy population structure I ignored any labels defined by sample site, as I used three different unsupervised clustering methods, Neighbor-Net, PCA and STRUCTURE, to detect population sub-structure.

Top level: Major unconnected geographical regions

Samples from the Caroni, Oropouche and Northern drainages of Trinidad, and from Venezuela were separated into four distinct clusters by Neighbor-Net (Figure 2.2), PCA (Figure 2.3A) and STRUCTURE analyses (Figure 2.4A). However, there was no unique clustering of samples from South West Trinidad, a region that was represented by two sampling sites only, namely Palm Drive Felicity (33) and Pitch Lake (34); numbers in parentheses refer to Table 2.1. Whereas fish from Palm Drive Felicity (33) which is close to the Caroni, clustered within the Caroni drainage, Pitch Lake (34) individuals from the far South West appeared as a distinct group in the Neighbor-Net (Figure 2.1). PCA indicated that both samples are part of the Caroni drainage cluster (Figure 2.3A). STRUCTURE analysis revealed predominantly the same genetic signature for individuals from the same geographical region, with a few notable exceptions. Individuals from the Turure in the Oropouche drainage (18, 19) were found within the Caroni drainage cluster (Figure 2.2, 2.3A and 2.4A) (see below). In agreement with the ancestry proportions inferred by STRUCTURE, fish from the lower Yarra (10) stood out among all samples from the Northern drainages in that they appeared equidistant between Northern and Caroni drainage clusters in the Neighbor-

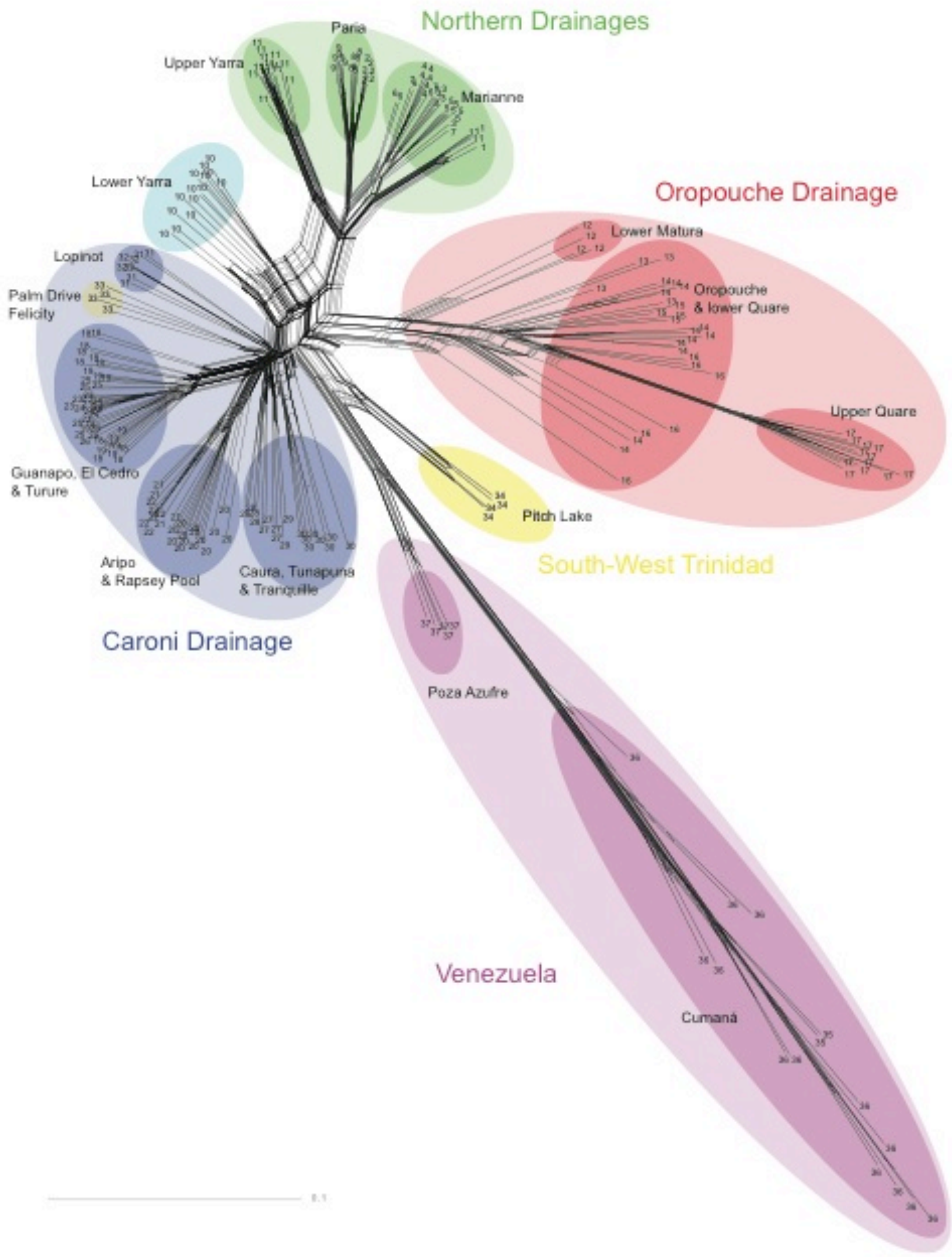


Figure 2.2: Phylogenetic network reconstructed with the method Neighbor-Net. The network is based on 866 SNP markers and shows all 239 individuals. Major geographical regions are color-coded as in Figure 2.1. Smaller circles within major geographical regions mark different river basins.

Net (Figure 2.2 and 2.4A). Samples from Poza Azufre (37) in East Venezuela shared genetic signatures with guppies from Cumaná, the Caroni, and the Northern drainages, but appeared to be more closely related to populations from the Caroni drainage in Trinidad than to populations from Cumaná or the Northern drainages (Figure 2.2 and 2.4A).

Second level: River basins

With all 239 individuals, likelihoods in Structure improved significantly with the number of clusters until k = 12 (Wilcoxon rank-sum test: P > 0.15 for k = 12 versus 13, ln P(k = 11) = -70069.80, ln P(k = 12) = -69229.70, ln P(k = 13) = -69584.47). For the majority of runs, clustering was very similar to that revealed by Neighbor-Net and to the one shown by k-means clustering at k = 12 using the first 11 PCs as input (Figure 2.2, 2.3A and 2.4B). The clustering clearly correlated with the second

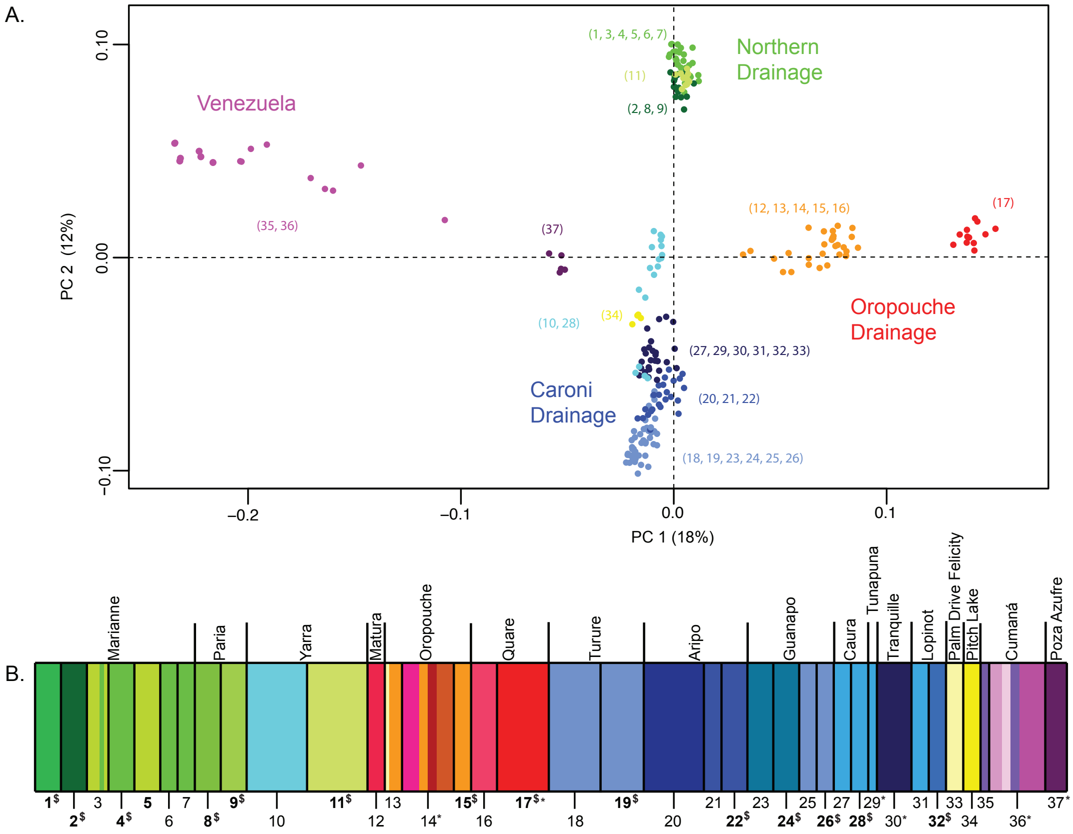


Figure 2.3: Principal component analysis (PCA) with k-means clustering. A) Plotting PC1 against PC2 revealed clustering reflecting four major geographical regions: Caroni drainage, Oropouche drainage, Northern drainages and Venezuela. Populations from South-West Trinidad (yellow) cluster with populations from the Caroni drainage. B) Clustering obtained by k-means using the all 28 significant PCs and k = 29. A vertical bar represents each individual. Vertical black lines indicate predefined regions, rivers and sample sites. Matching upper and lower samples are (10/11), (13/15), (16/17), (18/19), (20/22), (23/24), (25/26), (27/28), (31/32). Major geographical regions are color-coded as in Figure 2.1 and 2.2, with more color gradations discriminating sub-clusters in B. $^{\$}$Upper river samples, *laboratory reared populations.

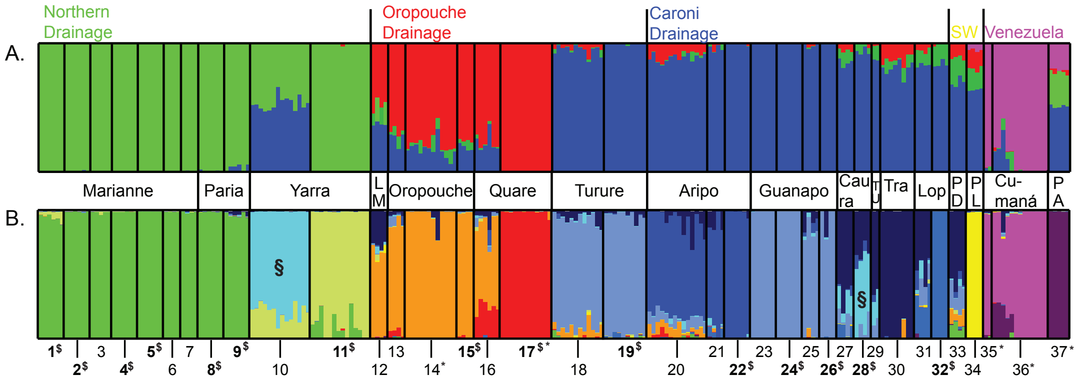


Figure 2.4: Analysis with Structure. Analysis based on 239 individuals using 806 SNP markers. Individuals are represented as vertical bars, horizontally partitioned into segments corresponding to their membership in genetic clusters indicated by colors. Vertical black lines indicate predefined regions, rivers and sample sites. Geographical regions are color-coded as in Figure 2.1 to 2.3. A) Clustering for *k* = 4 (ln P(*k* = 4) = -88159.53) B) Clustering for *k* = 12 (ln P(k = 12) = -69229.70). Abbreviations: SW = South-West Trinidad, Tu = Tunapuna, Tra = Tranquille, PL = Pitch Lake, PD = Palm Drive, LM = Lower Matura, PA = Poza Azufre. $^{\$}$Upper river samples, §undefined source population for admixture, *laboratory reared populations.

hierarchical level (river basins) of the sampling scheme (Table 2.1), although the three methods were applied without any *a priori* assumptions about populations.

Neighbor-Net (Figure 2.2) and STRUCTURE (Figure 2.4B) both grouped individuals from the Caroni drainage into clusters representing four of the rivers sampled, namely Aripo, Guanapo, Lopinot and Caura (Table 2.1). Samples from the Tranquille and Tunapuna fell into the Caura cluster. Clustering with *k*-means classified the Lopinot, which is represented by two sampling sites only (31, 32), as part of the Caura cluster. This observation reflects the close proximity of the Lopinot to the remaining rivers in this cluster (Figure 2.1). Guppies from the lower Aripo (20) showed evidence of genetic admixture with Oropouche fish (Figure 2.4B). Individuals from the upper Caura (28) showed admixture with a source population that did not appear to be included in the dataset, since there was no homogeneous cluster representing the source of the admixture. In addition, samples from the lower Yarra (10) located on the Northern slopes predominantly showed a genetic signature from apparently the same source population (Fig. 2.4B; marked with §).

At the second hierarchical level, all three analyses suggested that individuals from Pitch Lake (34) in South-West Trinidad constituted a separate cluster. Individuals from Palm Drive Felicity (33), which is also located in South-West Trinidad but much closer to the Caroni, remained in clusters representing the Caroni drainage, with contributions from clusters representing the Oropouche and Northern drainages (Figure 2.3C and 2.4B).

Within the Oropouche drainage I had samples from the Quare, Oropouche, Turure and Matura (Table 2.1). In agreement with the analysis at the first level (see above), all three methods showed the samples from the Turure (18, 19) as part of the cluster representing the Guanapo (23 – 26). This is explained by a known introduction from the lower Guanapo (25) into the upper Turure (19) about 50 years ago (Magurran, 2005). I observed that about 6% of the loci within lower Turure individuals (18) displayed alleles that were native to the Oropouche drainage population (Figure 2.4B), in agreement with previous studies (Becher and Magurran, 1999; Shaw, et al., 1992). The three methods revealed a clear separation between individuals from the upper Quare (17) and the remaining individuals from lower Quare (16), Oropouche (13, 14, 15) and lower Matura (12). Only the Neighbor-Net (Figure 2.2) classified the individuals from lower Matura as a separate subpopulation, whereas STRUCTURE and PCA suggested the lower Matura individuals formed part of the Oropouche cluster. In the Oropouche, I observed additional genetic signatures typical for Caroni drainage populations. Surprisingly, this admixture was not caused by individuals originating from the Guanapo and migrating downstream the Turure (Figure 2.4B), but seemed to originate from the Caura. This is remarkable, given that compared to the other Caroni drainage rivers, the Caura is relatively remote from the Oropouche drainage.

We included samples from the Yarra, Marianne and Paria, which are part of the Northern drainages in Trinidad (Figure 2.1). The Neighbor-Net as well as PCA indicated three separate clusters representing these rivers. Yet, individuals from a sampling site located within a tributary of the Marianne (2) were genetically similar to samples from the Paria (8, 9) (Figure 2.2 and 2.3A). Lower Yarra individuals (10) showed minor admixture from the upper Yarra (11), but major admixture from an unidentified source (see above).

The mainland of Eastern Venezuela was represented by two geographically well-separated sites, Poza Azufre (35) and Cumaná (36) (Figure 2.1). PCA, STRUCTURE and Neighbor-Net all suggested that these two samples were genetically very different (Figure 2.2, 2.3A and 2.4A).

Lowest level: Sample sites

Although likelihoods calculated by STRUCTURE did not improve significantly when more than 12 clusters were predefined, I detected 28 significant PCs by PCA, each having a p-value < 10^{-20}, suggesting at least 29 minor subpopulations within the dataset (Patterson, et al., 2006).

The *k*-means clustering revealed that the PCs described additional substructure within the Paria (2,8,9), Marianne (1,3-7), Oropouche (12-16), Aripo (20-22), Guanapo (18, 19, 23-26), Caura (27-32) and Cumaná (35-36) clusters (Figure 2.3B). These results suggested that each headwater sample site (Table 2.1; asterisks in Figure 2.3B) contained a subpopulation that had genetically diverged from communities in neighboring rivers. Samples from lower river ranges usually grouped with the corresponding upper river samples (Figure 2.3B), with three exceptions, namely samples of the lower Aripo (20), lower Quare (16) and lower Lopinot (31) were separate from their corresponding upper river populations (Figure 2.3B).

In addition to natural separation of sample sites, the maintenance of our laboratory fish stocks originating from the Oropouche drainage and Cumaná (see Material and Methods) may have imposed some artificial substructure as revealed by PCA. The clustering reproduced clearly the separation of samples in different families (Figure 2.3B). I therefore repeated the analysis without these laboratory-reared specimens and confirmed that these had not distorted the overall results (Figure 2.5). I also ran STRUCTURE on subsets corresponding to the five major geographical regions (Figure 2.6). PCA indicated even more refined sub-clustering within the Oropouche and Caroni drainages than STRUCTURE (Figure 2.3B and 2.5), reflecting plausibly the

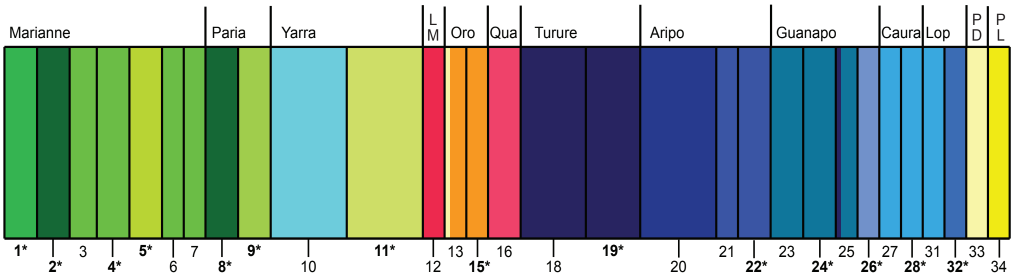


Figure 2.5: Principal component analysis (PCA) with k-means clustering. We excluded the 10 laboratory strains from the PCA in order to make sure that they did not distort our results. Clustering was obtained by *k*-means using the all 18 significant PCs and *k* = 19. A vertical bar represents each individual. Horizontal black lines indicate predefined regions, rivers and sample sites. Major geographical regions are color-coded as in Figure 2.3, with more color gradations discriminating sub-clusters. Abbreviations: PL = Pitch Lake, PD = Palm Drive, LM = Lower Matura. *Upper river samples

geographical structure. In addition, STRUCTURE indicated gene flow between neighboring downstream sites, in contrast to the more homogeneous appearance of upstream samples (Figure 2.6).

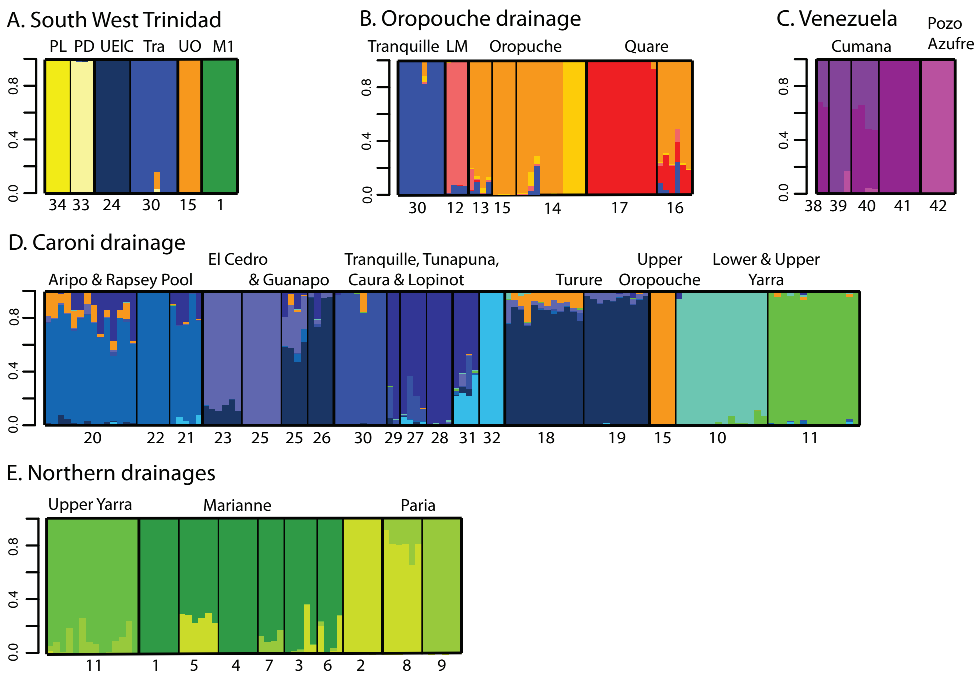


Figure 2.6: Detailed investigation of regional subpopulations. The 239 individuals were subdivided into five smaller sets representing the five major geographical regions. Populations that appeared as likely sources of admixture in the previous analysis (Figure 2.4) were included in every subset. Shown are the most likely STRUCTURE results for each subset (see Material and Methods). The analysis of subsets showed more refined sub clustering within region compared to the global analysis in Figure 2.4. Abbreviations: PL = Pitch Lake, PD = Palm Drive, UElC = Upper El Cedro, Tra = Tranquille, UO = Upper Oropouche, LM = Lower Matura, PA = Poza Azufre, ElC = El Cedro, Lop = Lopinot

2.2.2 Inference of population parameters

Population Divergence

A hierarchical analysis of molecular variance (AMOVA) revealed that 31.3% of the variance segregated among individuals within populations and 25.8% among sample sites within geographical regions (Table 2.2). However, the largest component of the global variation, 42.9%, was observed among regions (Table 2.2). PCA (Figure 2.3A) also provided information about the distribution of variation found within the dataset. It implied the highest variation between Cumaná and Quare samples, since these were placed at opposite ends of PC1, which harbors 18% of the variation (Figure 2.3A).

PC2 split samples from the Caroni and Northern drainages and explained 12% of the variation found in the dataset (Figure 2.3A).

AMOVA results on four subsets corresponding to the four major geographical regions indicated that the Caroni and Oropouche samples had the smallest values for variation between rivers (31.5 and 24.3%) and the largest values for variation within sample sites (56.0 and 60.2%, respectively). The highest value for variation between groups (46.8%) was observed between the two Venezuelan habitats, Cumaná and Poza Azufre, consistent with the Neighbor-Net analysis (Figure 2.2). Excluding the specimens reared in the laboratory did not alter the results significantly (Table 2.3).

Table 2.2: Analysis of molecular variance (AMOVA).

	Among groups	Among populations within groups	Within populations	Number of loci used
Hierarchical Level	1st	2nd	3rd	
Geographical regions	42.9***	25.8***	31.3***	573
Northern drainages	35.3**	30.1***	34.6***	624
Oropouche drainage	24.3^{NS}	15.5***	60.2***	656
Caroni drainage	31.5***	12.5***	56.0***	591
Venezuela	46.8^{NS}	16.6***	36.6***	686

NOTE: The major regions were the uppermost level, followed by river basins within these regions and the sample sites again at the lowest hierarchical level. ***P < 0.001, **P < 0.01, *P < 0.05, ^{NS} not significant.

Table 2.3: AMOVA without lab-reared specimens.

	Among groups	Among populations within groups	Within populations	Number of loci used
Hierarchical Level	1st	2nd	3rd	
Geographical regions	38.3***	27.1***	34.6***	573
Northern drainages	35.3***	30.1***	34.6***	624
Oropouche drainage	13.6***	5.2***	81.2***	656
Caroni drainage	34.8***	10.8***	54.4***	591

NOTE: The major regions were the uppermost level, followed by river basins within these regions and the sample sites again at the lowest hierarchical level. ***P < 0.001, **P < 0.01, *P < 0.05, ^{NS} not significant.

I found significant pairwise F_{ST}-values > 0 ($P < 0.05$) between almost all samples originating from different sites. F_{ST} values ranged from 0.036 [lower (18) and upper (19) Turure samples, Supplementary Table S2.1] to 0.908 [Paria (9) and upper

Lopinot (32), Supplementary Table S2.1]. Although genetic divergence seemed to be higher among populations from different major geographical regions there was also substantial allele frequency divergence among populations within regions. Allele frequency divergence was especially high among upper river populations, whereas lower river populations appeared to be less diverged from each other (Supplementary Table S2.1).

Heterozygosity

To test whether there was a significant association between heterozygosity and location within the river, I analyzed fish from high and low predation habitats separately and estimated the expected heterozygosity (H_e). H_e was significantly higher in lower river habitats (high predation) (mean $\mu = 0.12$, standard deviation $\sigma = 0.043$) than in the corresponding upland river (low predation) ($\mu= 0.06$, $\sigma = 0.037$; Wilcoxon rank-sum test: $P < 0.001$ (without laboratory strains); Figure 2.7).

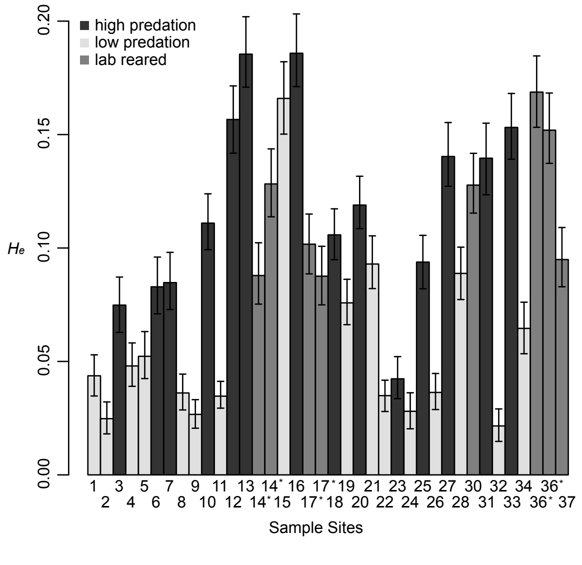


Figure 2.7: Summary of expected heterozygosity for specimens from different sample sites. We estimated the expected heterozygosity (H_e) for each population with $n > 3$. Additionally, 95% confidence intervals were indicated. Each sampling site can be categorized into either lower or upper river habitat (see Table 2.1). Laboratory reared populations are marked with asterisks. 14* Oro209 and Oro2; 17* Qua6 and QuaII_203; 36* CCELB and CCFR (see Table 2.1).

2.2.3 Regions under selection and candidate genes

The probability of detecting selection based on a significant F_{ST} greatly improves with population sample size. Although our sample sizes per site were rather small, defining populations by separate geographical regions effectively increased our sample size per population.

I scanned 310 SNPs that were polymorphic among individuals from the three separate geographical regions of Northern Trinidad, Caroni drainage (N=70), Oropouche drainage (N=42), Northern drainages (N=63), for signatures of directional selection using FDIST2 and BayesFst (Beaumont and Balding, 2004; Beaumont and Nichols, 1996). Ten runs of FDIST2 gave an average of 16.9 markers that were significantly scored to be under directional selection. This number is much higher than what one would expect by chance (3.1, $\alpha = 0.01$). In total, 31 different markers were significant in at least one run, with 12 SNPs (3.9%) with an F_{ST} value above the 99% confidence level reported in all ten runs (Table 2.8).

Table 2.8: Summary of the outlier loci detected by the Bayesian (BayesFst) and summary statistics (fdist2) methods.

FDIST2 Marker	BayesFst Marker	LG	Position (cM)
	0581^{*}	$1^{\S}$	18.476
0455^{*}	0455^{**}	$1^{\S}$	19.014
	1004^{**}	5	4.818
0417^{**}		5	6.932
0249^{*}	0249^{*}	5	22.17
	0232^{*}	6	2.813
0785^{**}	0785^{**}	6	22.365
0290^{**}	0290^{**}	7	1.52
0085^{**}	0085^{*}	$8^{\$}$	1.711
0999^{*}	0999^{*}	8	4.52
0026^{*}		$10^{\$}$	12.604
	0642^{*}	11	30.769
	0228^{*}	12	23.351
	0574^{*}	14	21.364
0628^{**}	0628^{**}	14	33.01
	0614^{*}	15	13.349
	1010^{*}	15	23.829
	0076^{**}	17	13.007
0893^{**}		19	8.96
0694^{**}		19	21.774
0280^{**}	0280^{**}	$20^{\S}$	0.735

NOTE: § Ornamental trait mapped to this region (Tripathi, et al., 2009), $^{\$}$Markers are linked to ESTs, *P < 0.01 (FDIST2) / P < 0.1 (BayesFst), **P < 0.005 (FDIST2) / P < 0.05 (BayesFst).

BayesFst reported on average 18 SNPs per run. In total, this approach detected 22 SNPs that were significant in at least one run, with 17 SNPs (5.5%) found in all ten runs. Eight markers were predicted with high probability by both methods to be under directional selection (Table 2.8). Of these, two were positioned near regions to which ornamental traits have been mapped (Tripathi, et al., 2009) and one was linked to an EST.

2.3 Discussion

I have carried out a comprehensive survey of genetic variation in wild guppy populations from Trinidad and Venezuela. The large SNP marker set I employed enabled me to apply modern approaches to the inference of population genetic history at the whole genome level. I found strong population substructure, which was highly correlated with the geographic features of the sampled area. My results are by large not in conflict with previous results obtained with other marker types. Thus, the SNP marker set is informative in many wild populations from different geographic regions in Trinidad and Venezuela and, therefore, a useful tool for analyzing demographic parameters and reconstructing lineage histories. In addition, I can both refine previous results, and provide new insights. Moreover, selection scans detected evidence for genetic differentiation among populations from different drainages in Northern Trinidad being not only due to genetic drift, but partly also to adaptive selection. Therefore, the data provided an important first step to future analyses of the genetics of adaptation in guppies.

Consequences of small sample sizes

To determine whether the marker set is useful for a large number of different populations, we wanted to sample very widely throughout the entire range of the species. This in turn necessitated that the sample sizes were limited per site (mean number of individuals 5.7). To reduce the impact of small sample sizes, I used individual based clustering methods, which fail to detect population substructure if sample sizes are too small and population differentiation is too weak (Patterson, et al., 2006). Estimates of population parameters such as F_{ST} and heterozygosity could still be affected by small sample sizes. A recent study by Reich *et al.* (2009) used similar sample sizes as this study to reconstruct human population history in India. They suggested a formula that has asymptotically minimal variance for estimation of F_{ST}.

Simulations have shown that one does not falsely encounter significant estimates of $F_{ST} > 0$ using sample sizes greater than four, if the expected F_{ST} between two populations equals zero (no genetic differentiation) (N = 2 to 20; 10,000 replicates per N tested with 1,000 permutations; see Chapter 3). If the expected $F_{ST} >> 0$, small sample size (N = 2 to 10) leads in general to a slight underestimation rather than overestimation of genetic differentiation (see Chapter 3). When estimating heterozygosity, one obtains a good estimate with moderate number of loci and sample sizes. However, variances in estimates decrease much faster with an increasing number of loci than with an increasing number of individuals (data not shown).

Given these observations and the fact that my results do not contradict previous studies leads me to the conclusion that I did not overestimate population substructure or heterozygosity due to small sample sizes. Instead, the large number of markers used provides me with a reliable genome wide picture of population substructure in wild guppies.

Past and present population structuring

I found clear genetic divergence among populations from the well-separated major geographic regions. Based on geological history and zoogeography of Trinidad, the "two arcs" hypothesis proposed that guppy populations from the Caroni and Oropouche basins had different origins on the South American mainland and have been separated since at least 600,000 years (Fajen and Breden, 1992). This stands in contrast to populations of the Caroni basin and the East Venezuela regions that may have been bridged until the end of Pleistocene (10,000 y b.p.; reviewed by (Magurran, 2005)). A neighbor-joining haplotype tree based on mtDNA had previously suggested a relatively close relationship between populations from the Northern slope, the Caroni basin and East Venezuela, including Poza Azufre, Cumaná, the Paria Peninsula, and the offshore island Isla Margarita. At the same time, it had suggested marked genetic divergence between these samples and those from the Oropouche drainage (Alexander, et al., 2006). These observations confirmed previous studies using allozymes that were also in support of the "two arcs" hypothesis.

My results are much more similar to what had been extrapolated from an analysis of seven microsatellites, which had suggested that populations from the Northern coast are highly differentiated from those in either the Caroni or Oropouche drainage (Suk and Neff, 2009). In addition, I propose shared ancestry for populations within the same geographical regions and that samples from the Caroni, Oropouche and Northern drainages of Trinidad are genetically about equally distant to each other. Shared ancestry is a novel finding for the populations from the three North flowing rivers for which this could previously not be detected (Alexander and Breden, 2004; Barson, et al., 2009; Carvalho, et al., 1991; Suk and Neff, 2009).

Populations from Poza Azufre and Cumaná were also included in my analysis. Based on morphological and behavioral traits, Alexander & Breden (2004) have proposed that the Cumaná guppy is highly differentiated from guppies originating from Poza Azufre, even though they had found little mitochondrial sequence variation. Based on nuclear markers, I suggest that there is strong genome wide genetic differentiation. Moreover, guppies from Poza Azufre appeared to be genetically more similar to Trinidadian than to Cumaná guppies, especially to those from the Caroni drainage (Figure 2.2, 2.3A and 2.4A). The Cumaná guppy exhibited the greatest genetic distance from all other populations studied. A caveat is that the genetic divergence of the Cumaná guppy might be overestimated due to a biased design that aimed at discrimination between Quare and Cumaná for the purpose of mapping crosses (Tripathi, et al., 2009). Therefore, some of the alleles occur only in either the Quare or the Cumaná populations, the two regions of maximal geographic distance in my study. Excluding the alleles that were specific to either Cumaná or Quare samples, decreased the genetic distance between the two populations, but did not alter the topology of the clustering (Figure 2.8). The distance between the Cumaná and Trinidadian guppies was most strongly affected, suggesting that there are more Cumaná- than Quare-specific alleles in our sample.

In conclusion, I propose based on genome wide SNPs that the populations from the three Northern drainages in Trinidad have been separated from both the Caroni and the Oropouche populations for as long as these have been sequestered from one another. The data supported genetically the notion that the Cumaná guppy may represent a case of incipient speciation (Alexander and Breden, 2004).

Previous studies focusing on single drainages found considerable differentiation among populations from different sites attributing it to waterfall barriers and geographic distance (Barson, et al., 2009); (Crispo, et al., 2006). In addition, Crispo *et al.* (2006) found that differences in predation or habitat features within the Marianne drainage did not influence genetic divergence or gene flow, while Barson *et al.* (2009) detected that lowland populations within the Caroni drainage experience gene flow

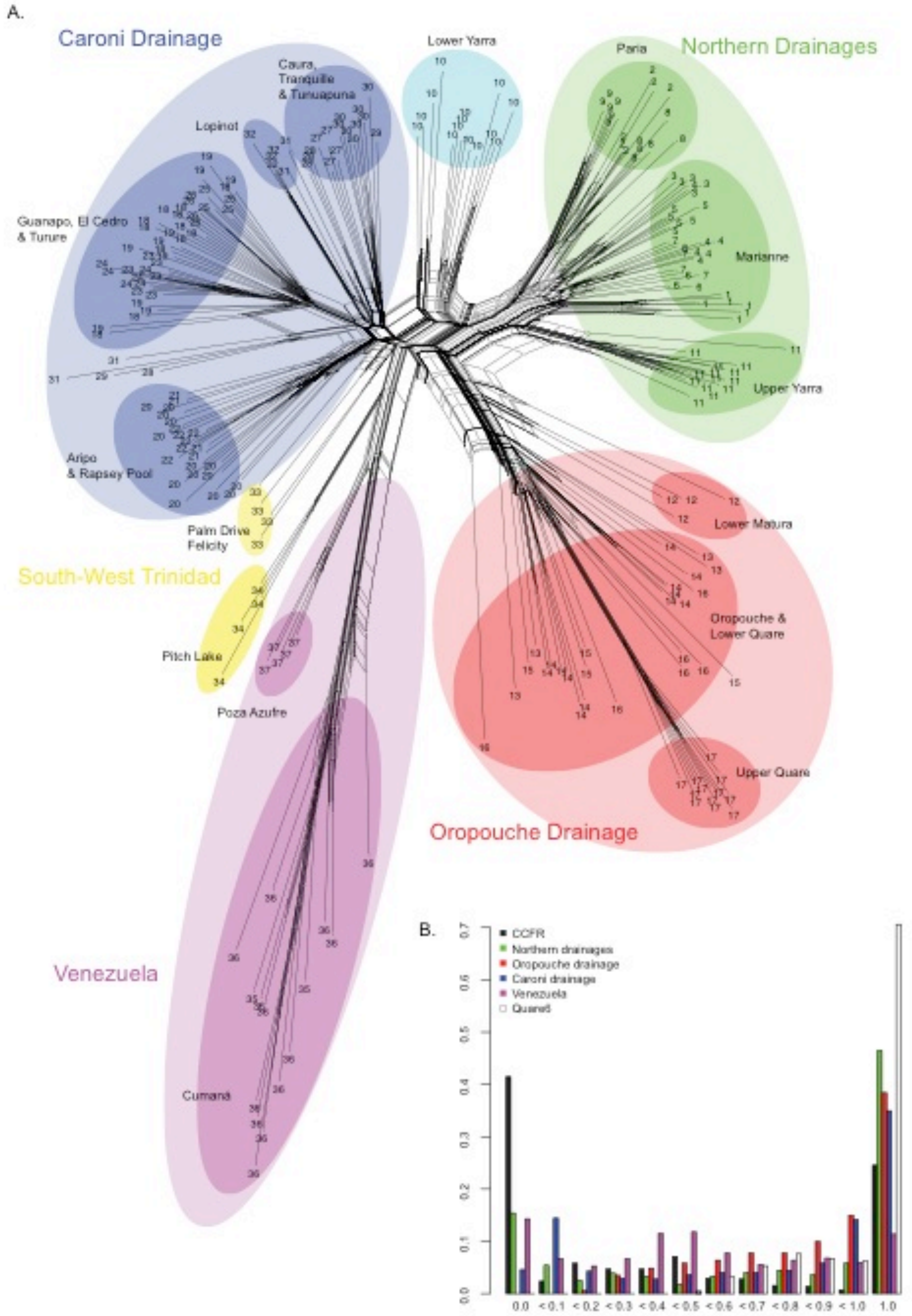


Figure 2.8: Influence of potential ascertainment bias. A. SNPs linked to alleles that were specific to samples from either Cumaná or upper Quare were excluded before reconstruction of a network using SPLITSTREE4. The clustering did not change compared to the one using all SNPs (Figure 2.2), but genetic distances between the Cumaná and remaining samples as well as between the upper Quare and the remaining samples decreased. B. Site frequency spectra using the major allele in upper Quare samples as reference. Allele frequencies of the reference were calculated for populations included in the ascertainment panel (Cumaná and upper Quare) and for samples within major regions.

from other lowland populations, with evidence for ongoing gene flow from headwater to lower tributaries. The pairwise F_{ST} values I estimated confirmed strong genetic differentiation between populations from different sites within the same drainage, especially between different headwater populations (Supplementary Table S2.1). I suggest that genetic differentiation is mainly shaped by geographic separation, which is evident from the hierarchical pattern found with the individual based clustering methods. In addition, matching upper and lower populations can be found in the same cluster (Figure 2.3B). This indicates that genetic drift is mainly responsible for the genetic differentiation, even if predator-driven selection is strong as proposed by the rapid adaptation detected in introduction experiments.

Admixture caused by gene flow among different downstream habitats could be visualized for the first time with STRUCTURE (Figure 2.4B). I found significant differences in heterozygosity between headwater and downstream populations supporting the notion of lower river populations being sinks receiving net gene flow (Barson, et al., 2009; Carvalho, et al., 1991; Shaw, et al., 1991). Since Barson et al. (2009) found little differences in effective population sizes (N_e) between upland and lowland populations within the Caroni drainage, we do not believe that N_e explains the detected difference in genetic diversity. However, if the headwaters were colonized by small numbers of individuals from the downstream regions, genetic diversity would also be reduced.

Signatures of admixture among drainages caused by introduction and natural gene flow

My analysis confirms that the guppies introduced by Haskins have largely replaced the original lower Turure population (Becher and Magurran, 1999; Magurran, 2005; Suk and Neff, 2009). Although there is no evidence that they have invaded the Oropouche, STRUCTURE revealed the presence of admixed alleles in Oropouche populations that appeared to resemble the Tranquille rather than the Turure (Figure 2.4B). Given the geographic distance between the two locations, it is unlikely that this admixture has a natural cause, but also no artificial introduction has been documented. There is a small genetic signature of Oropouche alleles within samples from the lower Aripo. Suk & Neff (2009) proposed based on seven microsatellites that Aripo populations are more similar to those in the Oropouche than to those in the Caroni drainage. This could be due to previously undetected admixture that might have been caused by natural gene flow, since the Quare and Aripo catchment areas are less than 70 m apart during the wet season (Magurran, 2005).

I found strong evidence for another admixture event between the lower Yarra (Northern drainages) and the Caroni drainage. Whereas to our knowledge no artificial introduction has been documented there, floods could have occasionally bridged the watershed between the Caroni basin and the Northwest slopes, resulting in some gene flow. This might have occurred between North-flowing rivers West of the lower Yarra, which originate close to the Western tributaries of the Caroni drainage (John Endler, personal communication). Since guppies can tolerate moderate salinities (Briggs, 1984), allowing for periodic colonization between river mouths, migration from guppies between adjacent river mouths could be possible (Carvalho, et al., 1991). Human interference cannot be excluded either. Since a linkage plot inferred by STRUCTURE (Supplementary Figure S2.1) revealed that the Caroni signature in the lower Yarra is evenly distributed across the genome, this admixture might be rather ancient. Another explanation is the occurrence of several independent colonization events of the Northern drainages, the Caroni drainage and the rivers further West to the lower Yarra, creating genetically diverged populations in these regions. To scrutinize either hypothesis, samples from rivers to the West of the Yarra and Tunapuna (see Figure 2.1) will have to be genotyped.

A surprising admixture pattern was observed between rivers within the Northern drainages of Trinidad. Individuals from sample site 2 (Table 2.1) in the Marianne River were genetically more similar to individuals from Paria. Sample site 2 is located in the Petite Marianne, a tributary that is separated from the main Marianne River (sample site M11 in (Crispo, et al., 2006)) by a high barrier waterfall, which effectively prevents upstream migration. On the other hand, the tributary has its origin near Paria tributaries, and it may be connected to the Paria catchment area during major floods. Therefore, it is likely that Paria rather than Marianne guppies have populated the Petite Marianne.

Signatures of selection in Trinidadian guppy populations

The combined results of two different F_{ST} outlier methods applied to the populations from three major geographical regions in Trinidad suggested between 3.5 and 6.5% of the SNP markers being under directional selection. This proportion lies in the range reported in other species using similar methods: 2.6 to 5.5% in humans (e.g. in (Akey, et al., 2002), 1.4 - 3.2 % in lake whitefish ecotypes (Campbell and Bernatchez, 2004), 2.6 - 3.3% in Norway spruce (Achere, et al., 2005), 9.5% in salmon (Vasemagi, et al., 2005), 1.3 to 3.6% in common frog (Bonin, et al., 2006), 5.5% in white spruce (Namroud, et al., 2008) and 5% in stickleback (Mäkinen, et al., 2008).

Two markers predicted to be under directional selection, 0581 and 0280, mapped to genomic regions previously scored by QTL mapping as candidates contributing to ornamental traits (Tripathi, et al., 2009). Only marker 0280 was scored as highly significant by both methods. The quantitative trait that mapped to the same region on chromosome 20 as marker 0280 is the area of a prominent orange spot on the central trunk (Tripathi, et al., 2009). Intriguingly, it has been reported that individuals from the Northern drainages exhibit more orange color compared to the remaining Trinidadian guppies, and this has been linked to a relatively red insensitive visual system of abundant guppy predators, prawns of the species *Macrobrachium* (Endler, 1983; Endler, 1991).

The association of some outlier alleles with QTL certainly needs scrutiny because QTL mapping for ornamental traits was at low resolution and did not include specimens from the Northern drainages of Trinidad. Furthermore, the strong historical subdivision between the three geographical regions may have produced false positives. Nevertheless these promising results suggest that genes under selection in geographically separated populations may be identified by F_{ST} outlier methods. Larger sample sizes will be required to identify genes under selection in contrasting habitats within the same river, but low/high predation contrast are replicated many times yielding great power to identify genomic regions under selection.

3. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using large panel of SNP markers

Studies on wild populations give important insights into population dynamics leading to genetic differentiation. One important goal of population genetic studies is to estimate the amount of genetic differentiation among populations in order to draw conclusions on the demographic history. A common measure for the degree of genetic differentiation is the fixation index F_{ST}, first defined by Wright (1951).

Until recently, most studies on wild population of non-reference species used moderately large numbers of samples per population (>20), but only a small number of genetic markers (< 20), preferentially microsatellites, for which more than two alleles can often be distinguished. Studies on human populations were among the first using thousands of markers, with single nucleotide polymorphisms (SNPs) as markers of choice. SNPs are typically the most abundant sequence variants in genomes. Their distribution throughout the entire genome at high density, combined with well-established models for handling mutation rates and error rates, and inexpensive methods for high throughput genotyping make them appealing for population genetic studies (Morin, et al., 2004). However, SNP assays are often designed using small panels incorporating only a fraction of populations and individuals that are later genotyped for these SNPs. Consequently, common polymorphisms are more likely detected than rare variants skewing allele frequencies to higher values (Rosenblum and Novembre, 2007). Additionally, because individual SNP assays are expensive to develop, studies on non-reference organisms, and particularly those on wild populations, are relatively rare (Narum, et al., 2008; Rosenblum and Novembre, 2007; Seddon, et al., 2005; Willing, et al., 2010). New methods incorporating next generation sequencing make it now possible to develop thousands of SNP assays with less bias and at a fraction of previous costs, also in non-reference organisms (Tautz, et al., 2010). It is commonly believed that large sample sizes (n > 20) are required to yield reliable estimates of differentiation (Holsinger and Weir, 2009). However, the question arises whether the large increase in the number of available genetic markers reduces the required sample sizes in order to get reliable estimates of F_{ST}. Reducing the sample size per population would make it possible to analyze a larger number of different populations at the same cost, and it offers an important advantage in conservation genetic studies on rare organisms.

The study described in this Chapter was conceived by Cock van Oosterhout, Christine Dreyer and me and is under preparation to be published (Willing et al., in preparation). I devised the experimental design, implemented the software used to conduct the simulations in order to examine whether the estimation of genetic differentiation measured by F_{ST} becomes inflated with small sample sizes and performed all the analyses described (see also Contributions). I concentrated the study on three different estimators. The first one was proposed by Wright (1951), which by definition lies between zero (no genetic differentiation) and one (population have gone to fixation for different alleles). However, Wright assumed infinite sample sizes in his definition, but population size is finite in real datasets. The absence of negative F_{ST} values in Wright's (1951) definition can lead to an overestimation of F_{ST}, particularly when the populations are only weakly or not differentiated. In such cases, estimates one fold higher than the actual differentiation cannot be compensated by estimates one fold smaller, because these would than be negative, and therefore, this estimator was expected to be upwardly biased. Cockerham and Weir (1984) proposed an unbiased estimator that can also have negative F_{ST} estimates and that has been widely used (Holsinger and Weir, 2009). Reich et al. recently published a population genetics study on human populations using very small sample sizes ($n = 4 - 6$). They proposed a new unbiased estimator for bi-allelic SNP data. Here I compared all three estimators for their performance on the same bi-allelic data set. This study addressed the following four questions. First, what is the type I error rate, i.e. falsely detecting genetic differentiation in a panmictic population, and what is the effect of small sample size? Second, does a small sample size result in an overestimation of the F_{ST} in cases where populations are genetically differentiated? Third, if estimates are biased by small sample sizes, can an increasing number of loci genotyped compensate for this bias? Fourth, what is the effect of ascertainment bias on the F_{ST} statistics, in particular, do deviations from the normal allele frequency distribution towards common or rare alleles lead to a bias in F_{ST} estimates?

3.1 Material & Methods

3.1.1 Data generation

I simulated an ancestral population with 1,000 individuals (50% males and 50% females, sex assigned randomly) and 21,000 bi-allelic loci. The genotypes at the loci were generated by randomly drawing from eight allele frequency classes (0.1,0.2,….,0.9). The two starting populations consisted of the same 1,000 individuals, which were genotyped at 10,000 loci randomly taken from the 21,000 loci of the ancestral population. I assumed an isolated island model (i.e. no migration between the two populations after separation). Genetic drift was simulated for a certain number (*t*) of generations according to the Wright-Fisher model without mutations (which is appropriate for SNPs, which arise at much lower rates than microsatellite variants). Consequently, the population sizes were kept constant, the generations were non-overlapping and the frequencies in the next generation were a binomial random sample based on the frequencies in the current generation. Random dioeciously mating was simulated by randomly drawing one female and one male with replacement. Thus all males and females were equally likely to be chosen and could mate multiple times, and the draw was independent of the number of times an individual has been chosen before. The two individuals drawn became the parents of one member of the next generation. Since I assumed all loci to be completely unlinked, I simulated gametogenesis by simply selecting at random one allele from each parent. This process was repeated until all members of the next generation have been created. I simulated different degrees of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2). The number *t* of generations was determined by $F_{ST} = 1-(1-\frac{1}{2n})^t$ with *n* equals the number of mating individuals (effective population size, N_e = 1,000) and *t* equals the number of generations needed to achieve the required amount of differentiation (Morin, et al., 2009).

3.1.2 Estimators tested

The first and simplest estimator, $F_{ST}{}^{W}$, was introduced by Wright in 1951. For one allele at locus *k*,

$$F_{ST}{}^{[k]} = \frac{s^2}{\bar{p}(1-\bar{p})}$$

where s^2 is the observed variance of allele frequencies p_i in the sampled populations i (i = 1, …, r), $s^2 = \sum_i (p_i - \overline{p})^2 / (r-1)$ and $\overline{p}$ is the mean allele frequency over all populations. The estimate of $F_{ST}{}^W$ for multiple loci is calculated by taking the mean across k loci.

$$\hat{F}_{ST} = \frac{\sum_k F_{ST}^{[k]}}{k}$$

This estimator has a theoretical range between zero and one and is known to overestimate the level of genetic differentiation especially at low values (Weir and Cockerham, 1984).

The second estimator tested, $F_{ST}{}^{W\&C}$, is probably one of the most widely used estimators. It was proposed by Weir & Cockerham (1984), who showed that it provides a nearly unbiased estimate of $\mathrm{F_{ST}}$ at moderate population sample size (n=15, 20 and 25) and small number of loci (k=10). The estimates can also have negative values which do not have a biological meaning (Weir, 1996), but they can compensate for overestimates especially at low levels of genetic differentiation. At a single locus k, $F_{ST}{}^{W\&C}$ is defined as

$$\hat{F}_{ST}{}^{[k]} = \frac{\hat{N}^{[k]}}{\hat{D}^{[k]}}$$

where

$$\hat{N}^{[k]} = s^2 - \frac{1}{2n-1}\left[\overline{p}(1-\overline{p}) - \frac{r-1}{r}s^2 - \frac{\overline{h}}{4}\right]$$

$$\hat{D}^{[k]} = \overline{p}(1-\overline{p}) + \frac{s^2}{r}$$

Here, s^2 is the observed variance of allele frequencies, n is the number of individuals per population, $\overline{p}$ is the mean allele frequency over all populations, r is the number of sampled populations and $\overline{h}$ is the mean observed heterozygosity. The overall estimate from all k loci is derived by

$$\hat{F}_{ST} = \frac{\sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}}.$$

Recently, Reich and colleagues (2009) proposed a new unbiased estimator, $F_{ST}{}^R$, for bi-allelic loci and pairwise population comparison. In their study they used a very high number of loci, but small sample sizes per population. Therefore, we decided to test this estimator as well. Again $F_{ST}{}^{R}$ is calculated as follows

$$\hat{F}_{ST}{}^{[k]} = \frac{\hat{N}^{[k]}}{\hat{D}^{[k]}}$$

$$\hat{N}^{[k]} = E((\frac{u}{2s} - \frac{v}{2t})^2) - \frac{E(\hat{h}_1)}{s} - \frac{E(\hat{h}_2)}{t}$$

$$\hat{D}^{[k]} = \hat{N}^{[k]} + E(\hat{h}_1) + E(\hat{h}_2),$$

where *u* is the allele count for population 1, *v* is the allele count for population 2, *t* and *s* are the total number of individuals for population 1 and 2, respectively. The parameter $\hat{h}_i$ is an unbiased estimate of the expected heterozygosity. An estimate over many loci is given by

$$\hat{F}_{ST}{}^{[k]} = \frac{\hat{N}^{[k]}}{\hat{D}^{[k]}}.$$

3.1.3 Statistical analysis

After *t* generations of random mating among 1,000 individuals, 10,000 of the 21,000 loci were randomly chosen to test the F_{ST} estimates. In order to test the influence of ascertainment bias in marker design, I generated three different datasets. The first set contained loci with equally distributed allele frequencies, the second set contained only loci with minimum allele frequency MAF > 0.25, because SNP marker sets are often biased in the direction of more common polymorphisms. However, I also generated a dataset of the other extreme containing only markers with MAF ≤ 0.25.

I used sample sizes of 2, 4, 6, 10, 20 and 50 individuals. For each sample size I sampled 10, 20, … , 100, 200, … , 1000, 2000, …, 5000 loci. For each number of individuals and genotyped loci I sampled from each population 1,000 times. I took the average F_{ST} estimate and derived the 95% confidence interval. I implemented a custom Java program to perform the simulations and estimations of F_{ST}.

3.2 Results

After *t* generations of random mating, I estimated F_{ST} on the complete simulated dataset comprising two populations with 1,000 individuals each that were genotyped at 21,000 loci. All estimators tended to give a slightly higher value than the theoretically expected F_{ST} (Table 3.1). The reason is that there is variance in the offspring number around the Poisson distribution, which slightly inflates the observed F_{ST} compared to the theoretically expected value.

Table 3.1: Estimated F_{ST} values on complete dataset

expected F_{ST}	*t*	F_{ST}^{W}	$F_{ST}^{W\&C}$	F_{ST}^{R}
0	0	0	-5.00E-04	-5.00E-04
0.0104	21	0.0107	0.0102	0.0102
0.0503	103	0.0542	0.0546	0.0547
0.1003	211	0.1073	0.1097	0.1096
0.2	447	0.2022	0.2068	0.2068
0.4001	1,022	0.4134	0.4024	0.4023

3.2.1 Estimates on SNP set with unbiased allele frequency distribution

I tested the influence of increasing the sample size on the estimate by taking 2, 4, 6, 10, 20 and 50 individuals from each population at different levels of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). Figure 3.1 shows an example were the number of loci is fixed at k = 100 and k = 1,000 at varying sample sizes. Figure 3.2 depicts an example were the number of individuals is fixed at n = 4 and n = 20 at increasing number of loci genotyped. Since combining all the different parameters (sample size, number of loci and level of genetic differentiation) resulted in a large number of estimates, I chose these four combinations in order to illustrate my general findings (all estimates can be downloaded as supplementary file). F_{ST}^{W} severely overestimated genetic differentiation in small sample sizes (n = 2 – 6) (e.g. Figure 3.1). Moreover, since this estimator cannot have negative values, the 95% CIs excluded zero implying significant genetic differentiation even if there is none. Also with moderate sample sizes (n = 10 – 50), F_{ST}^{W} slightly overestimates the level of genetic differentiation. Since these observations were consistent for all datasets, I will in the following concentrate on the behavior of the two other estimators.

The estimators $F_{ST}^{W\&C}$ and F_{ST}^{R} gave on average similar, fairly good estimates at all sample sizes (Figure 3.1). Importantly, both estimators did not indicate genetic differentiation when there was not any (Figure 3.1). However, if genetic differentiation was moderate or large ($F_{ST} \geq 0.1$) the $F_{ST}^{W\&C}$ estimator tended to slightly overestimate genetic differentiation with small sample sizes ($n \leq 6$), whereas the estimator F_{ST}^{R} showed the same average estimate irrespective of samples size. However, with increasing sample sizes the size of 95% CIs decreased. The 95% CIs

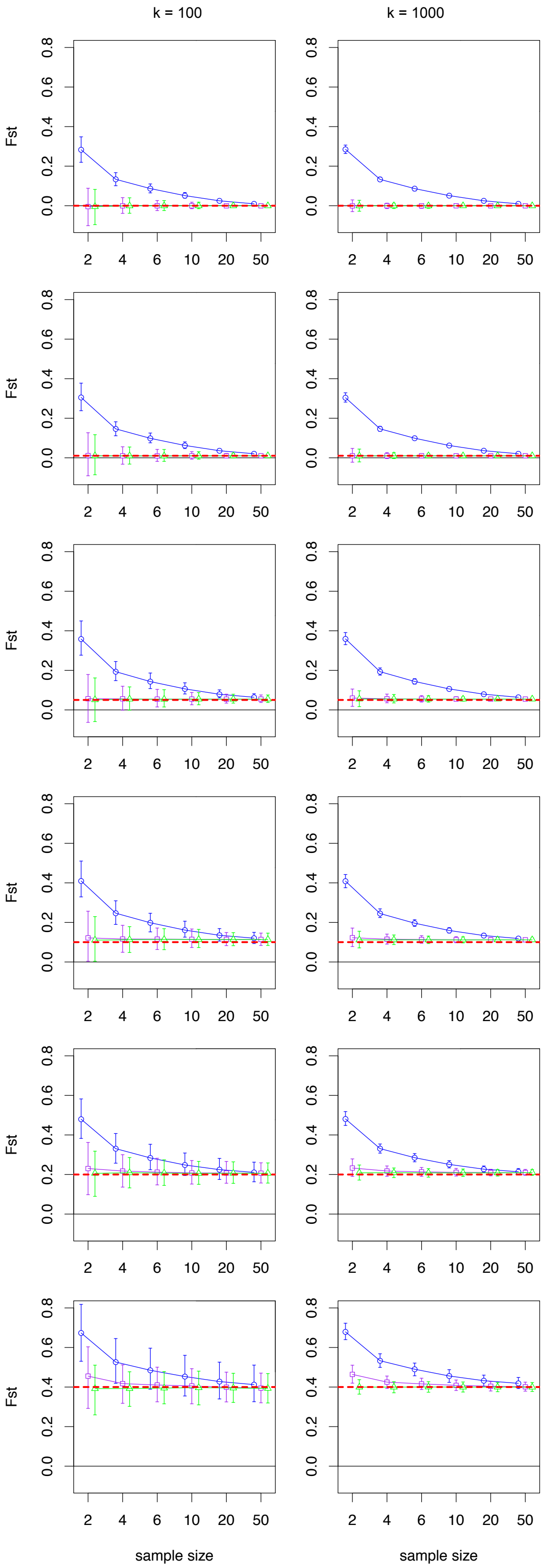


Figure 3.1: Effect of increasing sample sizes. Results are shown for the simulated data with equally distributed allele frequencies. Number of loci is fixed at k = 100 (left column) and k = 1,000 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: ${F_{ST}}^{W}$ (blue circles), ${F_{ST}}^{W\&C}$ (purple squares) and ${F_{ST}}^{R}$ (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

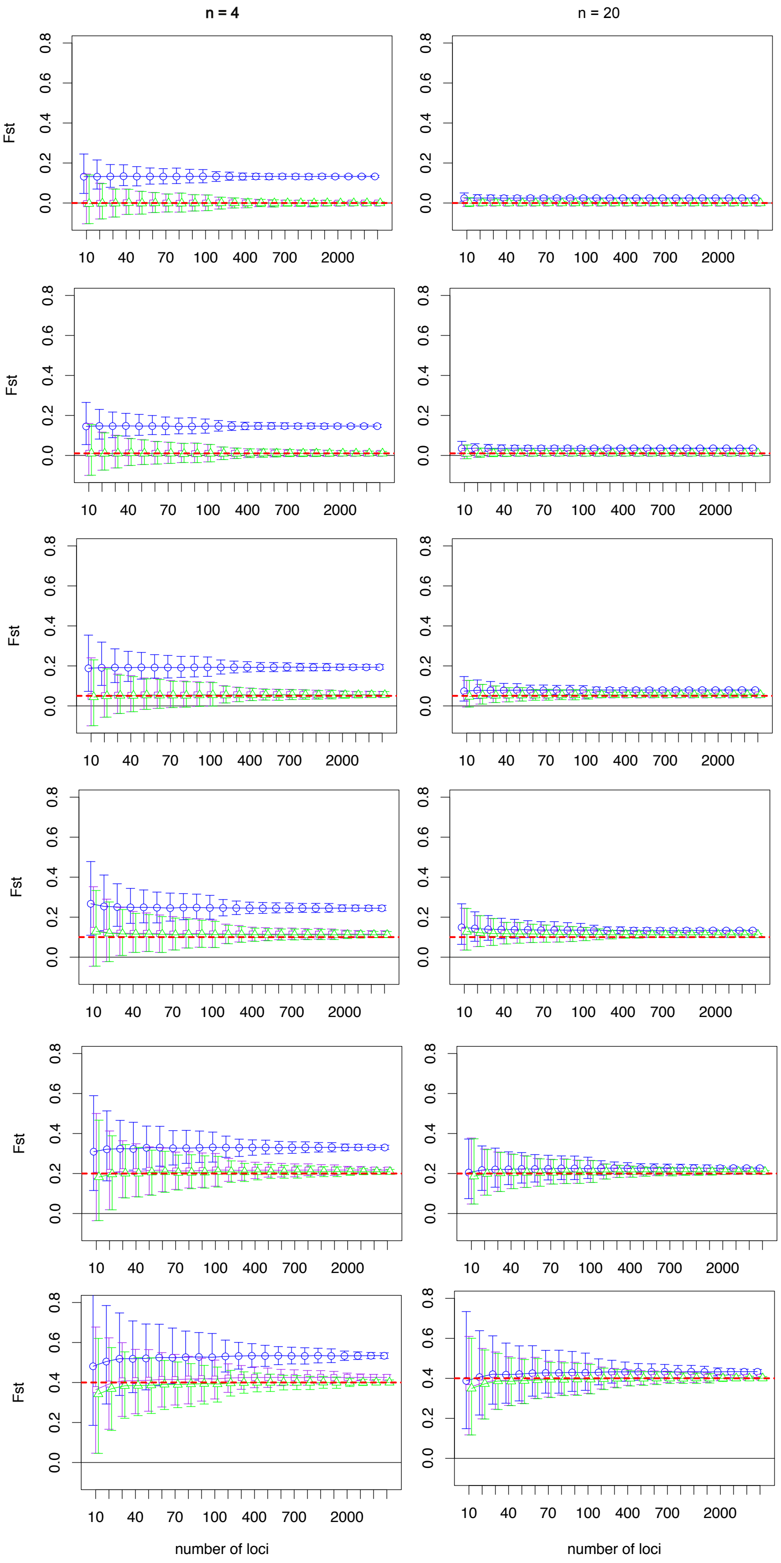


Figure 3.2: Effect of increasing the number of markers. Results are shown for the simulated data with equally distributed allele frequencies. Number of individuals is fixed at n = 4 (left column) and n = 20 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: $F_{ST}{}^{W}$ (blue circles), $F_{ST}{}^{W\&C}$ (purple squares) and $F_{ST}{}^{R}$ (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

were large and included zero at low genetic differentiation (F_{ST} < 0.05), when using a small sample size (*n* = 2 – 6) and a moderate number of loci (Figure 3.1, *k* = 100). Increasing the number of loci had no impact on the average estimate of F_{ST} (Figure 3.1, k = 1,000; Figure 3.2), but it did significantly reduce the 95% CIs. This effect was similar to the reduction in 95%CI caused by increasing the sample sizes (Figure 3.2). With 50 individuals per population and 1,000 loci one can detect genetic differentiation as small as 0.001 (average = 0.0011, 95% CI = [1.12E-04, 0.0022], Supplementary Table S3.1). Genetic differentiation as small as 0.01 can already be detected with *n* = 4 and *k* = 3000 (average = 0.0102, 95% CI = [0.0014, 0.0212], Supplementary Table S3.2).

3.2.2 Influence of differences in sample sizes

Next, I considered the impact of differences in samples sizes on the F_{ST} estimates. For this I kept the sample size taken from population 1 fixed at n_1 = 4 and varied the sample size taken from population 2. At low genetic differentiation ($F_{ST} \leq 0.05$) differences between sample sizes did not have an impact on the average estimates of F_{ST} of either estimator ($F_{ST}^{W\&C}$ and F_{ST}^{R}). If genetic differentiation was moderate to high ($F_{ST} \geq 0.1$), the $F_{ST}^{W\&C}$ overestimated genetic differentiation when the sample size of population 2 was small ($n_2 \leq 6$), but it gave an underestimation in cases where the sample taken from population 2 was large (n_2 = 50) (Supplementary Figure S3.1). The F_{ST}^{R} estimator gave on average the same estimate of F_{ST} independently of the differences between sample sizes. Furthermore the estimates were always very close to the expected level of genetic differentiation (Supplementary Figure S3.1 and S3.2). Therefore, I would recommend F_{ST}^{R}, if sample sizes differ among the populations analyzed. However, the magnitude of the 95% CI did not decrease with increasing the sample size in only one population of either estimator leading to the conclusion that the accuracy of an estimate depends on the smaller sample size taken (Supplementary Figure S3.1 and S3.2).

3.2.3 Estimates on SNP sets with biased allele frequency distributions

In order to test the influence of biases of allele frequency distribution in the analyzed marker set, I generated two datasets with 10,000 loci each, where in one set MAF > 0.25 and in the other set MAF ≤ 0.25. The simulations show that a bias towards common polymorphisms in the marker set leads to overestimation of genetic differentiation, whereas a bias towards rare polymorphisms leads to underestimation. These biases were observed in both the F_{ST}^{R} estimator as well as the $\mathrm{F_{ST}}^{W\&C}$ estimator, and they could neither be compensated by increasing the sample size nor by increasing the number of loci genotyped (Supplementary Figure S3.4 to S3.6). This suggests there will be a systematic overestimation of genetic differentiation due to ascertainment bias when the marker loci in the panel are developed based on the screening of a small number ($n < 4$) individuals.

3.3 Discussion

Our simulations show that even when sample sizes are small ($n = 2, 4, 6$), accurate and unbiased estimates of $\mathrm{F_{ST}}$ can still be obtained when a large number of bi-allelic markers such as SNPs are used, as long as the appropriate $\mathrm{F_{ST}}$ estimator is chosen. The original $\mathrm{F_{ST}}^{\mathrm{W}}$ estimator severely overestimates the level of genetic differentiation when using small sample sizes. Since this estimator cannot have negative values, these results were expected for values of $\mathrm{F_{ST}} < 0.5$, because overestimates that are higher than 0.5 over the actual $\mathrm{F_{ST}}$ cannot be compensated by a negative value at another locus. The two other estimators I tested showed similar performance on large sample sizes ($n \geq 20$), but the estimator proposed by Reich et al. (2009) showed better performance in cases where sample sizes were small ($n \leq 6$). Our simulations suggest that genetic differentiation is not falsely detected due to small sample sizes using these unbiased estimators. Furthermore, I showed that increasing the number of genetic markers has no impact on the mean $\mathrm{F_{ST}}$ estimates, but that it considerably reduces the 95% CIs. However, the accuracy of a pairwise estimate depends on the smaller sample size taken from one of the populations.

A previous study suggested that increasing the sample size might be more beneficial than increasing the number of markers genotyped (Morin, et al., 2009). However, that study tested a rather small number of SNPs (k < 100). Our study suggests that using a large number of markers (>500) increases significantly the power of detecting genetic differentiation even if using a small sample size. For example, pairwise genetic differentiation as small as 0.01 can be detected by taking a sample of only four individuals from each population when genotyped at 3,000 loci. This finding has important implications for studies on endangered species or those with small population size. By developing markers using next generation sequencing tools, conservation genetic studies can obtain the same statistical power in some of their population genetic analysis as studies performed on model organisms. However, testing datasets with biased allele frequencies (MAF ≤ 0.25 and MAF > 0.25) has shown that ascertainment bias has a severe effect on the estimation of F_{ST} rather than the sample size taken. Baird and colleagues (2008) proposed a method that has been proven particularly useful to develop a large number of genetic markers in non-reference organisms with less ascertainment bias (Hohenlohe, et al., 2011; Pfender, et al., 2011). Using multiplex strategies, samples taken from different populations in the wild can now be sequenced and genotyped in one lane of Illumina GAIIX sequencer (Elshire, et al., 2011). Therefore, it is now possible to analyze genetic differentiation from a large number of populations at low cost. Our simulations have shown that the cost of these analyses can be even reduced further by using only a small number of individuals per population.

4. Paired-end RAD-seq for de-novo assembly and marker design without available reference

In Chapter 2 I showed that for a large number of interesting questions a high number of genetic markers equally distributed over the genome is already very informative, even without a complete genome sequence. Therefore, new methods incorporating NGS for cost-effective high-throughput marker development are of great interest. Baird and colleagues (2008) developed a protocol for high throughput sequencing of restriction site associated DNA (RAD) tags using the Illumina platform (RAD-seq). It has the advantage that only a reduced representation of the genome is sequenced leading to deep sequence coverage of fragments near a specific type of restriction site. They showed that single end (SE) sequencing of RAD tags could be used for rapid marker development in stickleback for which a reference genome is available. Since then, SE RAD-seq has become a popular tool in next generation population genetics (Davey and Blaxter, 2010; Emerson, et al., 2010; Hohenlohe, et al., 2011; Hohenlohe, et al., 2010; Pfender, et al., 2011). In addition, Illumina PE sequencing could extend the sequence information on each side of the restriction sites (Baird, et al., 2008; Davey and Blaxter, 2010). Because each RAD can provide a unique genomic sequence tag that can be characterized without its immediate genomic context, the first reads may be aligned to each other, building subsets that are associated to one restriction site each. As a strategy for obtaining longer sequence tags, I exploited the fact that random mechanical shearing leads to a family of staggered second reads that can be assembled to longer subsets associated to the RE site defined by the first read cluster. This strategy subdivides the assembly problem into a high number of less complex local assemblies. This Chapter describes a study, published in Bioinformatics in 2011 (Willing et al., 2011), about PE RAD-seq data from two very diverged guppy populations, namely Quare and Cumaná, which have been previously used to generate a genetic linkage map (Tripathi, et al., 2009). Christine Dreyer, Margarete Hoffmann and I conceived the experimental design of the RAD-seq libraries. Margarete Hoffmann prepared the RAD-seq Illumina libraries analyzed. Juliane Klein implemented the assembly tool LOCASopt that was conceived and designed by me and considers the special needs for the assembly of PE RAD-seq data.

I conceived, designed and implemented the remaining tools described here and conducted all analyses (see also Contributions). I showed that my approach can generate *de-novo* 283,842 RAD tags that are 200 – 400 bp long and cover ~10% of the guppy genome. Furthermore, these tags can be used as reference to design thousands of new polymorphic markers useful for population genetic and mapping studies. All tools developed for the analysis are available as a package called RApiD.

4.1 Material & Methods

4.1.1 Creation and sequencing of the RAD library

The genomic RAD libraries were created as described by Baird and colleagues (Baird, et al., 2008). Briefly, genomic DNA pooled from six individuals each was digested with EcoRI (G'AATT,C, NEB). Pools represented Cumaná and Quare males and females and technical replicates of Quare males and Cumaná females were included (Table 4.1). Illumina P1 adaptors including a unique 12-bp multiplex identifier (MID) preceding the EcoRI site were added by ligation. All MIDs differed by at least seven bases and were therefore tolerant to up to three errors. After ligation of the P1 adaptors containing the different MIDs, the six DNA samples were pooled in proportionate amounts before shearing (Covaris) and addition of the P2 adaptor. A single library with an insert size range of 200 – 400 bp was prepared and sequenced from both ends with 100 bp read lengths in one lane of an Illumina GAIIX sequencer (Figure 4.1A).

Table 4.1: Sequence information and read counts for each 12 bp MID.

MID	Sequence	Sample	Million reads
1	ATGTGTCGCCAA	6 Quare males*	4.6
2	TCTGAGCGTACA	6 Quare males*	3.4
3	GATCTGAAGCTC	6 Quare females	0.015
4	CGACGATACTTG	6 Cumaná males	5.1
5	CTAGATGCTGAC	6 Cumaná females*	4.4
6	GACACCGTATGT	6 Cumaná females*	5.4

*technical replicates

4.1.2 De novo assembly of RAD tags

For quality control, I checked all first reads for presence of the partial 5 bp EcoRI motif (AATTC) following the 12 bp MID. Second, all reads containing uncalled

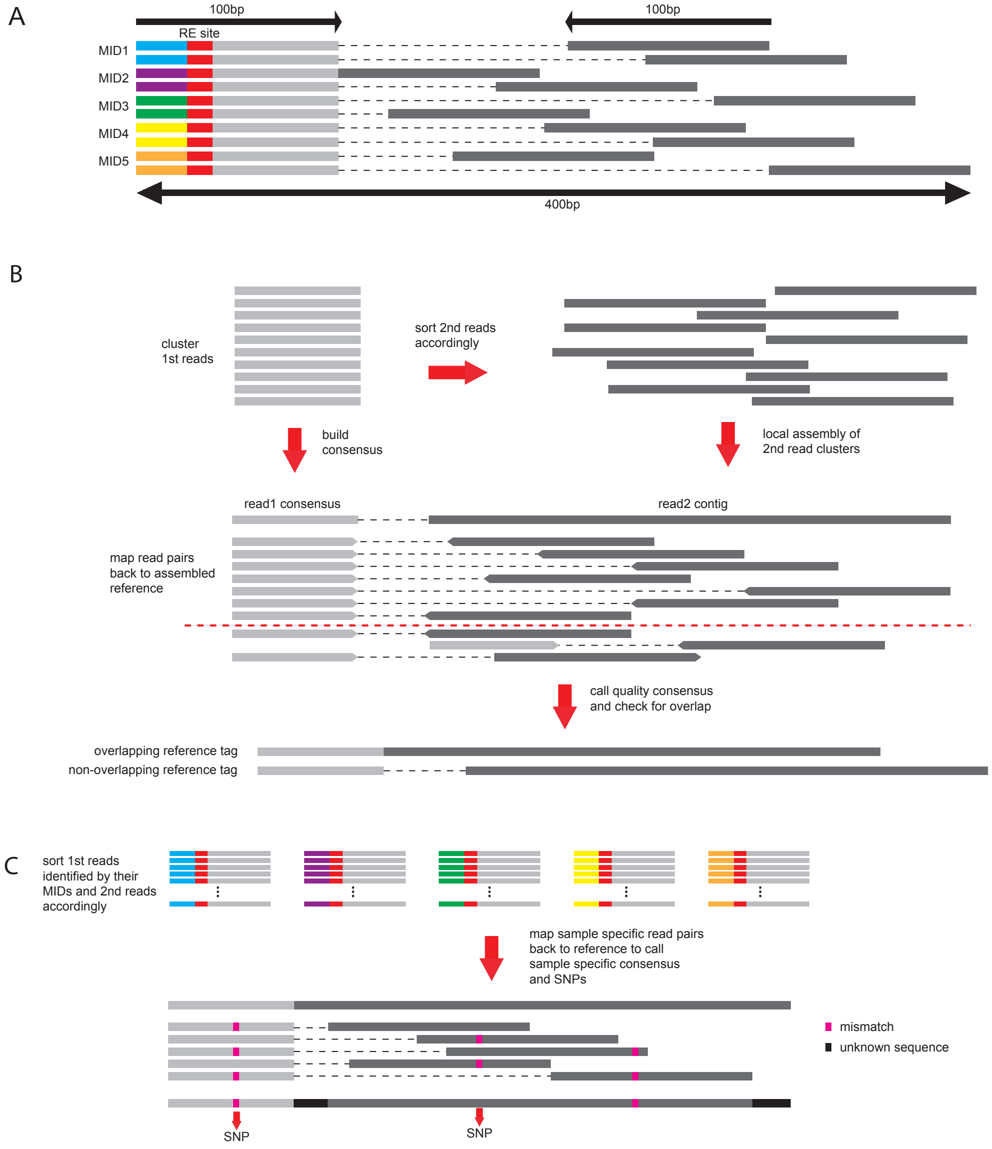


Figure 4.1: (A) RAD-seq output. Fragments are sheared randomly. By paired-end sequencing of fragments between 200 and 400 bp, the obtained 2nd reads are staggered and cover a range of 100 – 300 bp, whereas the 1st reads contain the MID and the restriction site and therefore start always at the same genomic position. (B) After removing the MID and the restriction site, the 1st reads can be aligned to each other (clustering). According to these clusters, the 2nd reads can also be sorted and assembled separately to a contig, which is then linked to the first read majority consensus. The tag pairs then serve as a reference to which all reads are mapped back and a majority consensus is called using only high quality bases from read pairs that mapped to the assembly fulfilling certain constrains (read pairs below red dashed line are discarded). After this step, all remaining tag pairs are checked whether they overlap. (C) All reads can be sorted according to their MID and sample specific consensus sequences and SNPs can be called by mapping the sorted reads back to the reference.

nucleotides were removed from the dataset. After removal of the MID and restriction site sequence the remaining first reads were grouped into pools representing the same RAD tag, using vmatch (www.vmatch.de), (Figure 4.1B). I allowed a maximal hamming distance of three within the same cluster. After clustering the first reads, the second reads could be sorted into groups accordingly that were assembled separately

(Figure 4.1B). Ideally, the local assembly of second reads in a cluster results in one contig indicating that they indeed originate from the same RAD tag. Mixed clusters of first reads could be caused by RE sites in repetitive regions. Such clusters might be resolved if the second reads were in a region outside the repeat and could be assembled into unique sequences. In such cases the assembly of the second reads resulted in more than one contig and allowed resolution of the mixed first reads accordingly. Every single cluster for each tag could have a different set of optimal assembly parameters, because of different repeat content and number of reads per cluster. E.g. if the coverage of a tag is low a smaller overlap length should be used for the assembly. I used the assembler LOCAS (Klein, et al., 2011) that uses an Overlap-Layout-Consensus approach to keep track of the overlaps among reads and is especially developed for low coverage data. LOCASopt is a wrapper that calls the assembler LOCAS with a different set of parameters in order to assemble the reads in a cluster several times under different conditions. Parameters that can be optimized are overlap length, percent of mismatches allowed in overlap, and seed size (see LOCAS Manual). LOCASopt keeps track of all the assemblies in order to choose the optimal one. I defined the optimal assembly as the one resulting in the smallest number of contigs and incorporating the largest fraction of available reads in a cluster. In order to test if optimizing each local assembly leads to better results, I assembled the data once with LOCASopt iterating over a large set of different parameter combinations, namely overlap = 21, 23, … 67, kmer = 13, 15, 17 and mismatch rate = 0.05, 0.07, 0.09. Additionally, I assembled the same set of clusters a second time with the parameters fixed at the values mostly used in the previous assembly (see Results).
After assembling, the 2nd read contigs are joined with the consensus of the corresponding first reads (Figure 4.1B). In order to generate a high quality reference I performed an additional quality control by mapping back all read pairs to the assembled tags and calling the majority consensus (see consensus and SNP calling) for each tag requiring a minimal quality of 20 (corresponding to a 0.01% chance that a base was wrongly called) and a minimal coverage of two per base. After that, uncalled nucleotides at the ends of the 2nd read contigs were removed. If there were uncalled nucleotides in the middle of a 2nd read contig, the contig was split up at these positions and the longest resulting substring remained as representative of this contig. Depending on the insert size of the library the 1st read consensus and 2nd read contig can be overlapping or non-overlapping (Figure 4.1B). Therefore, I checked for an

overlap between the two parts requiring a minimal overlap length of 10 bp and a maximal mismatch rate of 5%.

4.1.3 Consensus and SNP calling

After generating a comprehensive high quality reference, reads were sorted according to their MIDs and separately mapped back to the reference (Figure 4.1C). I used GenomeMapper (Schneeberger, et al., 2009) to map the reads back to the reference allowing up to five mismatches and no gaps. A mapped read pair has to pass several quality controls to be considered for consensus or SNP calling (see Figure 4.1B). Both reads in a pair have to map to the same contig in the right direction, with the start of the first read at the first position in a tag. A pair is only considered if at least one member uniquely maps to one contig in the reference. Furthermore, redundant read pair clones are removed in order to prevent false positive SNP calls that were caused by errors occurring during the amplification of the library. A read pair was considered to be a redundant clone, if the second read maps to the same position in a reference tag as another second read in a previous pair. After mapping the reads back the consensus base for each position in the reference was called by determining the major base at that position in the reads that could be mapped back. I used only bases with a minimal quality of 20 for consensus calling. Each consensus base got a quality value that was the average over the quality values of the bases used for consensus calling. If a position in the assembled reference was not covered during the consensus calling it was marked with a ‘N’ as uncalled nucleotide.

The search for polymorphic sites was done in a similar way as the consensus calling. A given site was considered polymorphic if the polymorphism occurred in at least a certain number of reads and if the site had a minimal coverage above threshold. In order to call a homozygous SNP, all reads must contain the same nucleotide that must be different from the reference. As in the consensus calling, only bases were considered that reached a certain quality threshold. The quality of a SNP is the average of the qualities of the single bases at the SNP position.

4.2. Results

4.2.1 Paired-end sequencing

In order to generate a dense set of RAD markers, we chose the restrictions enzyme EcoRI, which recognizes the palindromic 6-bp sequence G'AATT,C. The guppy genome size is nearly 1 Gb as estimated by flow cytometry (Manfred Schartl, personal communication). Based on sequenced BAC ends from a genomic library of the Cumaná guppy, the guppy genome was predicted to be relatively AT-rich (60%), close to the AT content of the EcoRI recognition site. For simplicity, I assumed that EcoRI sites occur close to the expected frequency of 1/4,096 bp, and that I have therefore an expected number of 500,000 RAD tags. To test the sequencing depth required as well as reproducibility of the results, six independently digested bulks of DNA from six individuals each were pooled, representing males and females of two different populations, and technical replicates (see Table 4.1). Paired-end (PE) sequencing with 101 bp read length of this pool on a single lane of an Illumina flowcell resulted in 23.4 million read pairs, of which 97% (22.6 million) contained the correct restriction site pattern (AATTC) at the beginning of the first read and no uncalled positions. Consequently, assuming 500,000 tags, each tag should be covered by ~46 read pairs on average. Figure 4.2 shows the base and quality score counts per site in each read. The base distribution over the first 17 bp in the first read nicely

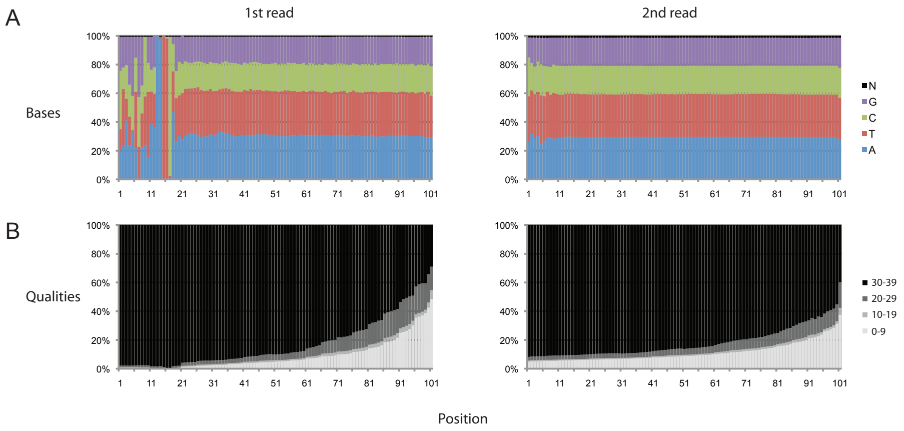


Figure 4.2: Base distribution and quality scores along the reads. (A) The number of bases per position in each read was determined. For read1, positions 1-12 contain the sample specific MID and at position 13-17 is the restriction site (clearly seen in the base counts at these position). Read2 is completely comprised of genomic bases. The base distribution at positions containing solely genomic regions reflects the expected distribution of ~60% AT. (B) Quality values were counted along the reads. As expected, quality values decrease to the end of the reads.

depicts the MIDs and the restriction site. However, at the first position after the restriction site G is significantly underrepresented, possibly as a consequence of genomic CG methylation inhibiting EcoRI. Yet after position 18 the distribution converges on the expected values (60% AT, 40% GC), which is seen over the entire second read. As expected, quality values of both reads decrease over their length (Bansal, et al., 2010), with a slightly faster decline in the first read, possibly caused by the unequal base distribution in the first 17 bp. For consensus and SNP calling the reads were sorted according to sample specific MIDs (Table 4.1). Differences between read counts for the different samples deviated less than a factor of 1.6 from each other, with one exception. This is within the range previously encountered when sequencing multiplexed samples (Craig, et al., 2008). Only approximately 15.000 reads encoded with MID3 were obtained, suggesting technical failure (Craig, et al., 2008).

4.2.2 Clustering and de novo assembly

All-against-all alignment of reads resulted in 451,981 first-read clusters with ~48 reads on average (range 2 to 66,393). For assembly, I considered only 297,147 (65.7%) clusters within a certain coverage range (5 – 184), in order to avoid highly repetitive regions. These clusters had an average size of 63 reads and included 18.9 million (81%) of the reads.

The second reads belonging to each first-read cluster were sorted and assembled separately to obtain a second-read contig for each cluster (Figure 4.1B). If the assembly of a cluster resulted in more than one contig or if not all the reads were used in the assembly, the first reads were sorted anew, according to the assembled contigs. I performed the assembly twice, once iterating over different parameter settings and once fixing the parameters at the values mostly used in the optimized assembly (overlap = 21, kmer = 13, mismatch rate = 0.05). The assembly with fixed parameters resulted in 503,748 contigs with an average length of 286 bp, representing 291,149 clusters and incorporating 76.6% of the reads. On average, 28 read pairs contributed to one RAD tag (Table 4.2). In the optimized assembly, 291,159 clusters were assembled resulting in 334,215 second-read contigs with an average length of 349 bp using 76.8% of the input reads. On average 43 read pairs contributed to one RAD tag, which is close to the 46 read pairs expected per tag (Table 4.2). Figure 4.3 shows that after optimizing the assembly the increase in the number of longer contigs was marginal, but most of the very short contigs may have been merged with longer contigs by choosing a different set of parameters. This notion is supported by the fact that significantly less clusters result in more than one contig in the optimized assembly (8.7% compared to 31.0%, Table 4.2). Consequently, optimizing the set of parameters for each local assembly led to less, but on average longer second read contigs with a higher number of reads used per contig. I therefore used these contigs for all following analyses.

Table 4.2. Results of the optimized versus the not optimized assembly.

	Fixed	Optimized
Clusters resulting in an assembly	291,149	291,159
Contigs	503,748	334,215
Clusters resulting in >1 contig	31.0%	8.7%
Reads used	75.6%	76.8%
Average contig length	286 bp	349 bp
Sum of all contigs	143.8 Mb	116.6 Mb

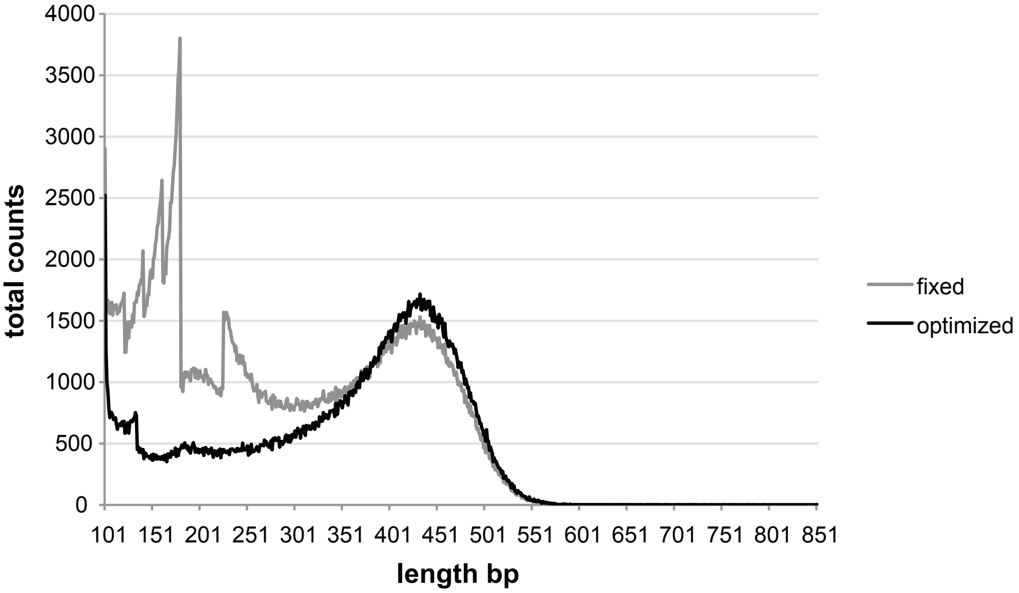


Figure 4.3: Length distribution of assembled contigs. The assembly of the 2nd read clusters was performed twice. Once with parameters fixed at certain values and once trying a large set of different assembly parameter combinations to find the optimal one for each cluster. The optimal assembly was defined as the one resulting in the least number of contigs, but incorporating the largest fraction of reads.

4.2.3 Quality control

Following the strategy detailed in Material and Methods, I found 283,842 contigs fulfilling the quality requirements, corresponding to ~57% of the tags expected. This is comparable to the number of EcoRI RAD tags found in stickleback (Baird, et al., 2008), where short (36 bp) reads were aligned to a reference genome. Of the assembled guppy RAD tags, 51.4% were overlapping with their corresponding first read consensus, over a length of 29 bp and with an average mismatch rate of 0.002%. For these tags I obtained on average 417 bp continuous sequence corresponding to 60 Mb in total. The second read contigs of the non-overlapping tags were on average 259 bp long. Taking the sequence information together from all overlapping and non-overlapping tags I obtained 108.2 Mb of sequence, corresponding to about one tenth of the guppy genome, close to the expectation.

In order to assess the quality of the de novo assembled reference, I predicted RAD markers of ≥150 bp length by digesting 6,165 sequenced Cumaná BAC ends with EcoRI *in silico*. These were used as queries in a Blastn search against our high quality reference. Of 1,112 predicted RAD markers, 862 (77.5%) matched (≤1e-100) our assembled reference. Of these 862 hits, 798 (92.6%) covered over 90% of the query or the subject sequence and included the restriction site at one end. These results show that our strategy led to a high quality reference of de-novo assembled RAD markers that can be further used for sample-specific consensus and SNP calling.

4.2.4 Sample specific consensus calling

After assembly and quality control, I sorted the reads according to their MIDs and mapped each batch on the reference in order to call sample-specific consensus sequences. Baird and colleagues (2008) used the presence or absence of a tag to identify it as polymorphic. The absence of a RAD tag in one sample is probably most often caused by a polymorphism in the associated restriction site. However, random sampling in the sequencing process can cause false positives. Therefore, Baird and colleagues (2008) scored only such markers as absent that were represented by at least eight reads in one sample and by none in the other sample. I tested whether this strategy also works with *de-novo* RAD tags by comparing the intersections between the different samples using different coverage cut-offs (1x, 6x, 10x) to assign a marker as polymorphic. The technical replicates provided the opportunity to estimate the false positive rate at the different coverage cut-offs. Table 4.3 shows how many tags I found per sample at different coverage cut-offs (diagonal) and the percentage of markers that could not be found in the intersection between the different samples and would therefore be scored as polymorphic. The false positive rate declines from greater than 1% with a minimum coverage of 1x to below 0.3% and 0.04% with minimum coverage of 6x and 10x, respectively. We see from Table 4.3 that the percentage of absent markers between the samples from the two different populations is much higher (>14% at all coverage thresholds) than the highest false positive rate.

Table 4.3: Pairwise comparison of missing RAD tags between the five sample pools using different coverage thresholds. Diagonal contains the total number of tags in the sample with the required coverage. Remaining entries give the percentage of RAD tags that can be found in sample *i* (rows) but are missing in sample *j* (columns).

MID	Minimum coverage	1	2	4	5	6
	1x	218,946	2.88	19.49	19.89	19.35
1	6x	174,735	0.26	16.81	17.07	16.75
	10x	137,679	0.04	15.11	15.31	15.06
	1x	1.89	216,720	19.53	19.93	19.37
2	6x	0.06	153,463	16.02	16.28	15.98
	10x	0.01	103,495	14.03	14.26	14.06
	1x	25.11	25.91	235,368	3.54	2.86
4	6x	20.70	21.17	185,295	0.73	0.60
	10x	18.13	18.46	148,954	0.34	0.32
	1x	25.17	25.97	3.15	234,404	1.80
5	6x	20.14	20.57	0.39	178,107	0.05
	10x	17.25	17.53	0.10	137,078	0.01
	1x	25.30	26.07	3.28	2.62	236,381
6	6x	20.97	21.47	0.50	0.15	190,100
	10x	18.27	18.60	0.12	0.02	155,298

I infer that a significant number of polymorphic markers were caused by sequence variation that changes restriction enzyme sites. At 10x coverage, less than 0.04% of these markers are false positives.

In guppies, sex is genetically determined and sex-linked inheritance and sex chromosome evolution are topics of general interest in this species (Lindholm and Breden, 2002). Sex is determined by male heterogamety (XY), but the master sex determining locus, which appears to be located at the distal end of the Y chromosome, has not yet been precisely mapped due to a lack of markers (Tripathi, et al., 2009). I inspected the *de-novo* assembled RAD tags for sex-specific markers. At 10x coverage, there were at least 2.5-fold more markers polymorphic (0.1% / 0.12% and 0.34% / 0.32%, Table 4.3) in the Cumaná female/male (MID 4 compared to 5 and 6, Table 4.3) contrasts, compared to the Cumaná female/female (0.02% and 0.01%, MID 5 and 6, Table 4.3) or Quare male/male (0.04% and 0.01%, MID 1 and 2, Table 4.3) contrasts, corresponding to ~149 female specific tags and ~477 male specific tags. Because 40% of these markers are expected to be false positives at 10x coverage, a higher coverage threshold should be used.

4.2.5 Distribution and fidelity of polymorphic sites

The distribution of polymorphic sites along the assembled RAD tags was analyzed by mapping all reads back to the assembled reference. A site was regarded as polymorphic if the polymorphism was covered by at least two reads and the coverage was at least six fold. SNPs were called with quality thresholds of either 20 or 30. Figure 4.4 shows that the coverage decreases significantly toward the end of the first read, with declining quality scores, as is typical for the end of the reads (Bansal, et al., 2010). Over the first 69 bp, SNPs are found with equal frequency at each position in the first read, but the number of SNPs significantly increases to the end of the first read even when using a quality threshold of 30. However, this might not only be caused by decreasing quality values at the end of the reads, but might be also due to more misalignments at the end of the reads. When I do not use the last 15 bp of each mapped read for SNP calling, I reduce the number of SNPs mainly at the proximal end of the second read part of the tag (red curve in Figure 4.4). Figure 4.4 also illustrates that the second read contigs have their maximal coverage around position ~270 bp and that the coverage decreases as expected toward both ends of the contigs. Furthermore, the likelihood to detect a SNP at a certain position in the second read part of a tag is positively correlated to the coverage. However, above a coverage of on average ~15 fold SNP detection does not seem to increase further, suggesting that such coverage is sufficient to detect the majority of alleles.

To determine the number of SNPs that could be confirmed in the intersection of technical replicates, I analyzed each sample separately. Based on the observations described above, I performed the sample specific SNP calling disregarding the last 15 bp of each read and considering only those positions in the reference having a coverage at least equal to a certain cut-off in all samples. For the Quare male

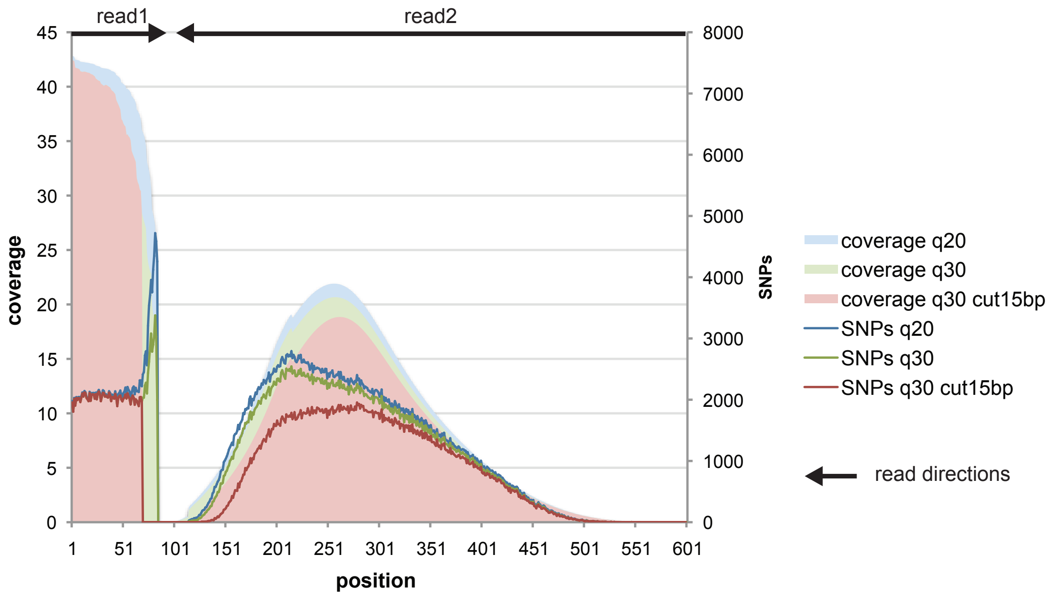


Figure 4.4: Distribution of polymorphic sites in assembled tags All read pairs were mapped back to the tag pairs in order to calculate the total coverage and the total number of polymorphisms per site as a function of the quality cutoff and read length used

replicates I used the Cumaná consensus as reference and for the Cumaná female replicates the Quare consensus as reference, in order to compare high fidelity rates for heterozygous as well as homozygous SNPs. At 6x coverage, 84% of the heterozygous SNPs within the Quare replicates, and 86% of heterozygous SNPs within the Cumaná replicates, could be found in the intersection. At 10x coverage, these numbers increased as expected slightly to 89% and 90%, respectively. In order to determine whether this applies to both parts of a tag, I examined the intersections of the first and second read part separately. I found that 87-91% of the SNPs detected in the first read lie in the intersection between the technical replicates, but only 78-80% of the heterozygous SNPs in the second read. Apart from the higher coverage in the first read, this could be also partly due to position dependent systematic errors in the base calling that are equally likely in each sample. Since the first reads in a RAD tag are completely overlapping, position dependent systematic errors can lead to false positive heterozygous SNPs that are shared among different multiplexed samples. The position dependent effect cannot occur in the second part of a tag because I did not consider read pair clones. However, homozygous SNPs differ from the heterozygous SNPs in their fidelity. At 6x coverage, I find over 97.1% of the SNPs in the intersection of the technical replicates, and this increases only to over 97.7% at 10x coverage. Moreover the intersections between the first and the second part differ by less than 1%. This indicates that the detection of homozygous SNPs between populations is highly reproducible with this method. Nevertheless, our approach also allows the detection of a high number of high fidelity heterozygous SNPs within populations at a specificity rate of >78% at comparatively low coverage.

4.2.6 Polymorphic markers within and between Quare and Cumaná populations

To determine the number of polymorphic markers within and between Quare and Cumaná specimens, I pooled the technical replicates to increase the coverage. At a minimum coverage of 6x, I found that 28.9% of the assembled 283,842 RAD markers are polymorphic between the two populations due to a polymorphism affecting the enzyme recognition site.

Including only those positions in the reference with at least 6x coverage in each population sample, I scored 302,693 polymorphic sites, of which 148,770 (49.1%) were homozygous SNPs differentiating the two populations, and 153,923 (50.9) sites contained SNPs that were heterozygous in at least one population. I found 116,861 (41.2%) tags containing at least one polymorphism of which 81,405 (28.7%) contained at least one homozygous SNP and 73,199 (25.9%) at least one heterozygous SNP indicating that some tags can be either scored for a homozygous or heterozygous SNP.

Using the complete set of 302,693 SNPs, I estimated the expected heterozygosity (H_e) (Excoffier, 2007) within each population and the genetic differentiation measured by F_{ST} (Reich, et al., 2009) between the two population samples. I found $H_e = 0.078$ and $H_e = 0.138$ within Quare and Cumaná samples, respectively. These values are similar to a previous study using genome wide SNP markers for population structure analysis (Willing, et al., 2010). The $\mathrm{F_{ST}}$ was estimated to be equal to 0.71. This is somewhat lower than what has been reported before (see Discussion).

4.3 Discussion

In this study I have demonstrated a method for the de-novo assembly and analysis of PE RAD-seq data. I was able to assemble ~10% of the guppy genome represented by 283,842 RAD tags of which ~50% were overlapping. This ratio could be significantly increased by either reducing the insert size of the library or by sequencing with longer read length. About 29% of the tags were polymorphic between the Quare and Cumaná populations due to a disruption in the EcoRI recognition site and about 41% of the tags contained at least one SNP site. Furthermore, some tags could be scored for a heterozygous and a homozygous SNP making the same tag a useful marker for different kinds of analyses (e.g. population genetics vs. genetic map). I have scored polymorphic sites using a newly developed approach, because the first read and error models developed for SNP calling in whole-genome sequencing data do not apply to RAD-seq data. As the first read of a specific tag starts at an invariant position, a SNP within the first read will always be at the same position. This is severely punished by some error models used for SNP calling, because sequencing errors at the same site are correlated (Li, et al., 2008). In addition, I do not expect a large number of insertions and deletions causing misalignments, because the reference is assembled with the reads that are also used for SNP calling. Moreover, repetitive sequences are removed by removing large first read clusters. These properties make the alignment problem fairly easy and eliminate the main sources of false positive SNPs in genomic data (Li, et al., 2008; Malhis and Jones, 2010). While my approach supports the use of other SNP calling algorithms using the assembled consensus tags as reference, I would advise to filter the mapping file used as input, following the criteria for informative reads defined in this study (see Material and Methods, Section 4.1.3). Estimated population parameters using the scored SNPs are similar to those previously reported. However, the estimated F_{ST} of 0.71 is somewhat lower, perhaps due to the less biased choice of markers compared to the previous work, which used markers designed for mapping crosses, with fixed SNPs between the two populations being preferred over segregating ones, inflating the estimation of genetic differentiation between the two populations (Willing, et al., 2010). Consequently, our approach will produce a high number of unbiased informative SNPs that are ideal for population genetic analyses. I found that 81,405 of the tags contain homozygous SNP between Cumaná and Quare populations. These would be potentially useful in generating a dense genetic map that would greatly aid a whole genome assembly. Furthermore, the paired-end RAD-seq contigs could be used as artificial long reads in a whole genome assembly, to overcome the problems of assembling an entire genome from short reads only. Moreover, one could use different restriction enzymes to generate an overlapping set of RAD-seq contigs. By counting the restriction sites of ten additional 6-cutter enzymes in our assembled data (unpublished data), I saw that

167,848 tags contain at least one of ten other restriction enzyme sites analyzed. Similar sequence complexity reduction approaches for aiding genome assemblies have been advocated before (e.g. Hyten, et al., 2010).

5. Sequencing of specific genomic regions contained in BACs with short reads

In Chapter 2, I have shown how a genome-wide set of genetic markers can give hints about genomic regions under selection by applying F_{ST} outlier methods and Chapter 4 introduced a new approach that makes it feasible to obtain these genome-wide marker sets. However, even if the marker density is high enough to pinpoint a genomic region under selection, the marker by itself does not give any hints about the genes located in its neighborhood. Although next generation sequencing has greatly facilitated the re-sequencing of reference organisms, *de-novo* sequencing and assembly of complex genomes remains a challenge using short reads only. In the described case above, however, one does not need to sequence the whole genome, but much smaller pieces containing genomic regions of interest. Bacterial Artificial Chromosomes (BAC) libraries contain the genome of an organisms fragmented into 100 to 350 Kb inserts. Wicker et al. (2006) did the first study on BAC shotgun sequencing using 454 sequencing. They sequenced four BACs containing parts of the barley genome in order to test whether or not 454 sequencing data is sufficient for de-novo assembly of complex genomes. They found that all coding fractions of the BACs were excellently covered in the assembly. The 454 read length at that time was 100 bp. Since then, this read length has greatly improved (> 500 bp) and sequencing of mate pairs has become available. Recently, more studies on BAC sequencing with 454 technology incorporating the new protocols have been published in barley (Sato, et al., 2011; Steuernagel, et al., 2009; Wicker, et al., 2006), rice (Rounsley, et al., 2009), melon (Gonzalez, et al., 2010) and salmon (Quinn, et al., 2008). In these studies not each BAC was sequenced separately but instead sequenced libraries with pooled DNA from 20 to 400 BACs. These BACs either originated from the same genomic location (Quinn, et al., 2008; Rounsley, et al., 2009; Sato, et al., 2011) or were non-overlapping (Steuernagel, et al., 2009). While the experimental designs where slightly different among these studies, they all showed that coding fractions could be excellently reconstructed *de-novo* and 454 sequencing of BACs has meanwhile become an established tool. However, Illumina sequencing has not been used so far for the sequencing of BAC pools, despite the much cheaper per base cost and the fact that the read length of up to 150 bp is now comparable to the initial 454 read length. I conceived and designed the study described in this chapter. Christine Dreyer, Verena Kottler and I selected the BACs for sequencing. Margarete Hoffman and Verena Kottler prepared the BAC DNA and Margarete Hoffmann made the Illumina libraries for sequencing. I performed all analyses described in this Chapter (see also Contributions). In the following, I will describe the *de-novo* assembly and analysis of Illumina sequencing data generated from a library containing the pooled DNA of eleven guppy BACs that were not barcoded prior to pooling (Table 5.1). Seven of these were of interest because they were associated to SNP markers scored to be under directional selection (see Section 2.2.3). Therefore, gene mining in these BACs could lead to candidate genes under selection.

5.1 Material and Methods

5.1.1 Sequencing and quality filtering

Eleven BACs linked to ten different mapped markers (Table 5.1) were obtained for sequencing from a guppy BAC library with an average insert size of 160 Kb, representing eight times coverage of the male Cumaná guppy genome (Tripathi, et al., 2009). Seven of these markers were of interest, because they were scored as being under directional selection (see Section 2.2.3), three were associated to a known QTL (Tripathi, et al., 2009) and one was chosen, because it was linked to a candidate gene for color patterning, namely *slc45a2*. The BAC library was screened by filter hybridization using specific probes for each BAC, if the parent clone was not a BAC end (Table 5.1). The DNA of each BAC was isolated using the QIAGEN large construct kit and then pooled to equal amounts in one sample. It is important that each BAC is present in the pool in sufficient quantity to be sequenced effectively and that no individual BAC dominates the pool. However, certain variability in the relative amounts of each BAC clone in the pool cannot be avoided. The library was prepared following the Illumina protocol for paired-end sequencing of genomic DNA. The insert size of the library was 200 bp and the sequencing was done on an Illumina GAII sequencer with 80 bp read length. The raw data was subjected to quality trimming using qualityTrimmer (Euler-sr package) (Chaisson and Pevzner, 2008) with a minimal base quality of six (-*minQual* 6) and all reads shorter than 50 bp after trimming were discarded (-*maxTrim* 30). Since the BACs were amplified within *E.*

coli using the BAC vector pIndigoBAC-5, a certain amount of contamination of *E. coli* and vector DNA could no be avoided. In order to remove these reads, I mapped all reads against the *E.coli* reference genome and pIndigoBAC-5 sequence using GenomeMapper (Schneeberger, et al., 2009), allowing a maximal edit distance of four (-E 4). After quality trimming and filtering, I built the intersection between the first and second read files in order to get pairing information. Those reads without partner after filtering were used as singles.

Table 5.1 Information to BACs sequenced with Illumina. *The parent clone that was used to develop the SNP marker, is the BAC end of the BAC sequenced

Parent clone	Accession (NCBI)	Marker	Guppy LG	Position (cM)	BAC sequenced	Type
Tra_Embryo_3_4_G01	ES371771	30	12	9.52	Bac15_I02	QTL
Blu_Testis_6_B19	ES380805	76	17	13.01	Bac25_P09	Selection
Tra_Liver_7_4_G02	ES377426	85	8	1.71	Bac28_O05	Selection
yaBac01_2_H02	FH888654	280	20	0.74	Bac01_O04*	Selection/QTL
yaBac01_3_D05	FH888700	290	7	1.52	Bac01_H09 Bac32_G02*	Selection
yaBac03_F10	FH889280	380	12	5.66	Bac02_F10*	QTL
yaBac03_I24	FH889361	396	8	4.49	Bac03_I24*	Selection
yaBac04_2_D06	FH889637	455	1	19.01	Bac04_G12*	Selection
yaBac33_1_A10	FH890456	581	1	18.48	Bac33_A19*	Selection/QTL
slc45a2	FJ236222	1075	12	-	Bac19_N24	Pigmentation gene

5.1.2 De-novo assembly

The *de-novo* assembly was done using the Velvet assembler (Zerbino and Birney, 2008), which is especially designed to build contigs and eventually scaffolds using short-read sequencing data. Velvet uses *de Bruijn* graphs as data structure to find overlaps between the reads. There are some critical parameters that can be set by the user and that substantially influence the outcome of an assembly. The probably most important one is the hash length (*k*) also known as *k-mer* length or word length. It corresponds to the length of substrings that are stored and compared in the construction of the *de-Bruijn* graph. In order to determine the optimal *k-mer* length, I first iterated over different values ranging from 31 to 69, increasing the value stepwise by two, because velvet only takes odd numbers as hash length to avoid palindromes. Another crucial parameter that can be given to the assembler is the expected *k-mer* coverage (*exp_cov*). It allows Velvet to determine which contigs correspond to unique regions of the genome and which contigs correspond to repeats. It can be determined using the following probabilistic formula

$$\exp\_\mathrm{cov} = \frac{(L-k+1)*bp\_\mathrm{cov}}{L},$$

where L is the read length and *bp_cov* stands for the base pair coverage that is determined by

$$bp\_\mathrm{cov} = \frac{R*L}{M},$$

where R is the number of reads and M the expected input in base pairs. The best assembly was ascertained by the highest N50. The N50 metric was here defined as the largest contig size at which half of the total size of the contigs is represented by contigs larger than the N50 value. Related to the expected *k-mer* coverage are the minimal coverage (*cov_cutoff*) and maximal coverage cutoff (*max_cov*) that can also be given by the user. Assembled regions with coverage below or above these thresholds are removed, because they have not enough support (coverage is too low) or are repetitive (high coverage is an indication for collapsed repeats). However, as already mentioned above, certain variability among DNA concentrations of the different BACs in the pool cannot be avoided. Therefore, it is likely that the coverage among contigs from different BACs is not equal. I therefore varied the coverage thresholds (t) after determining the optimal k in order to find the optimal cutoffs by setting

$$\mathrm{cov}\_cuttoff = \frac{1}{t}*\exp\_\mathrm{cov}$$

and

$$\max\_\mathrm{cov} = t*\exp\_\mathrm{cov}$$

with t = 5, 10, 15, 20, 25 and 30. The ends of each BAC were previously Sanger sequenced and could be used to identify each BAC insert boundary. For some BACs I had additional Sanger sequence information like parts of cDNAs, ESTs or genomic marker sequences that should be contained in the BAC inserts (Supplementary Table S5.1). In order to ascertain that contigs from all BACs were covered in the assembly, the best assembly was identified as the assembly in which the maximal number of Sanger sequences could be found by blasting (*e-value* < 1e-50).

5.1.3 Assigning the contigs to BACs and gene mining assuming synteny in other species

Since the contigs in our assembly were a mixture of the pooled BACs, I had to come up with a strategy to assign each contig to one BAC. Medaka (*Oryzias latipes*) and stickleback (*Gasterosteus aculeatus*) are the two most closely related fish species to guppies with sequenced genomes. By blasting (blastn) (Altschul, et al., 1990) the contigs against these reference sequences, I hoped to identify clusters of contigs aligning to a syntenic region and, therefore, belonging to the same BAC. I used GBrowse (Stein, et al., 2002) in order to visualize the blast results. The clusters corresponding to one BAC could be identified by contigs within a cluster having a significant hit to at least one of the respective Sanger sequences of a specific BAC (Supplementary Table S5.1). After assigning each contig to a BAC, I blasted (blastn) all contigs against NCBI non-redundant (nr) database for annotation. In addition, I blasted the contigs of each cluster separately against the stickleback and medaka reference genomes using the Ensembl Genome Browser (www.ensembl.org) in order to use the provided annotation of regions with the majority of significant hits (*e-value* < 1e-4).

5.2 Results

5.2.1 *De-novo* assembly using Velvet

Eleven BACs were pooled to equal amounts in one library without barcoding. Sequencing on a single lane of an Illumina GAII flowcell resulted in ~23.3 Mio. read pairs. After quality trimming ~19 Mio. first reads and ~8.2 Mio. second reads remained, indicating that the first reads were of better quality. The filtering step revealed that only ~2.3% of the reads were originating from *E.coli* and ~4.4% from the pIndigoBAC-5 Vector, leaving us with ~13.8 Mio. paired and ~11.6 Mio. single reads from the BAC inserts, which were on average 78 bp long. These reads were *de-novo* assembled using Velvet.

Since one BAC has an average length of ~160 Kb, the amount of base pairs sequenced was estimated to be ~1.76 Mb leading to an expected base pair coverage of ~1,125x on average. Iterating over different hash lengths showed that $k = 53$ led to the best assembly defined by the highest N50 value of 6,819 bp and maximal contig length of 38,458 bp (Table 5.2).

After finding the optimal hash length, I repeated the assembly with different values for the minimal (*-cov_cutoff*) and maximal (*-max_cov*) coverage cutoff keeping k = 53. In order to find out which assembly contained most likely contigs from all BACs, I blasted the available Sanger sequences against each assembly. At a coverage cutoff of t = 20, most Sanger sequences had a significant hit to at least one of the contigs in the assembly (*e-value* < 1e-50, Supplementary Table S5.1). Increasing t to 30 did not lead to additional significant hits of previously missing Sanger sequences. In addition, increasing t further than 20 did neither significantly improve the N50 value of 8,557 bp nor the maximum contig length of 38,458 bp (Table 5.3). However, the number of contigs was significantly reduced from 20,124 to 2,096 by providing coverage cutoffs (Table 5.2 and Table 5.3). I used the assembly done with k = 53 and t = 20 for further analysis.

5.2.2 Assigning contigs to BACs using syntenic regions

In order to assign contigs to a specific BAC I blasted all contigs against medaka and stickleback reference genome sequences. I used the search function in the locally installed GBrowse (Stein, et al., 2002), where I visualized the alignments of contigs to the reference genomes, to look for the contigs containing the Sanger sequences (Supplementary Table S5.1). If these contigs were located near a cluster of aligned contigs, I assumed that these were contigs from the corresponding BAC (e.g. Figure 5.1). Contigs that appeared in conserved order in the stickleback and medaka reference genome were believed to belong to the same BAC (e.g. Figure 5.1). With this strategy, I was able to identify a cluster for each BAC located chromosome 21 in medaka and the corresponding group XVI in stickleback that could not be identified by a BAC end. I labeled this cluster as Marker 0380, since the remaining BACs could be ascertained (see Discussion). In this way, I was able to assign 187 of 716 contigs larger than 500 bp to the BACs. The coverage among all contigs assigned to a BAC varied drastically between 20x and 1,241x with an average coverage of 203x and a standard deviation of σ = 238.5. However, coverage among contigs from the same BAC was much more similar (σ = 5 to 39, Table 5.4). The only exception was Marker 0085 with σ = 190, which is almost as high as the standard deviation of coverage

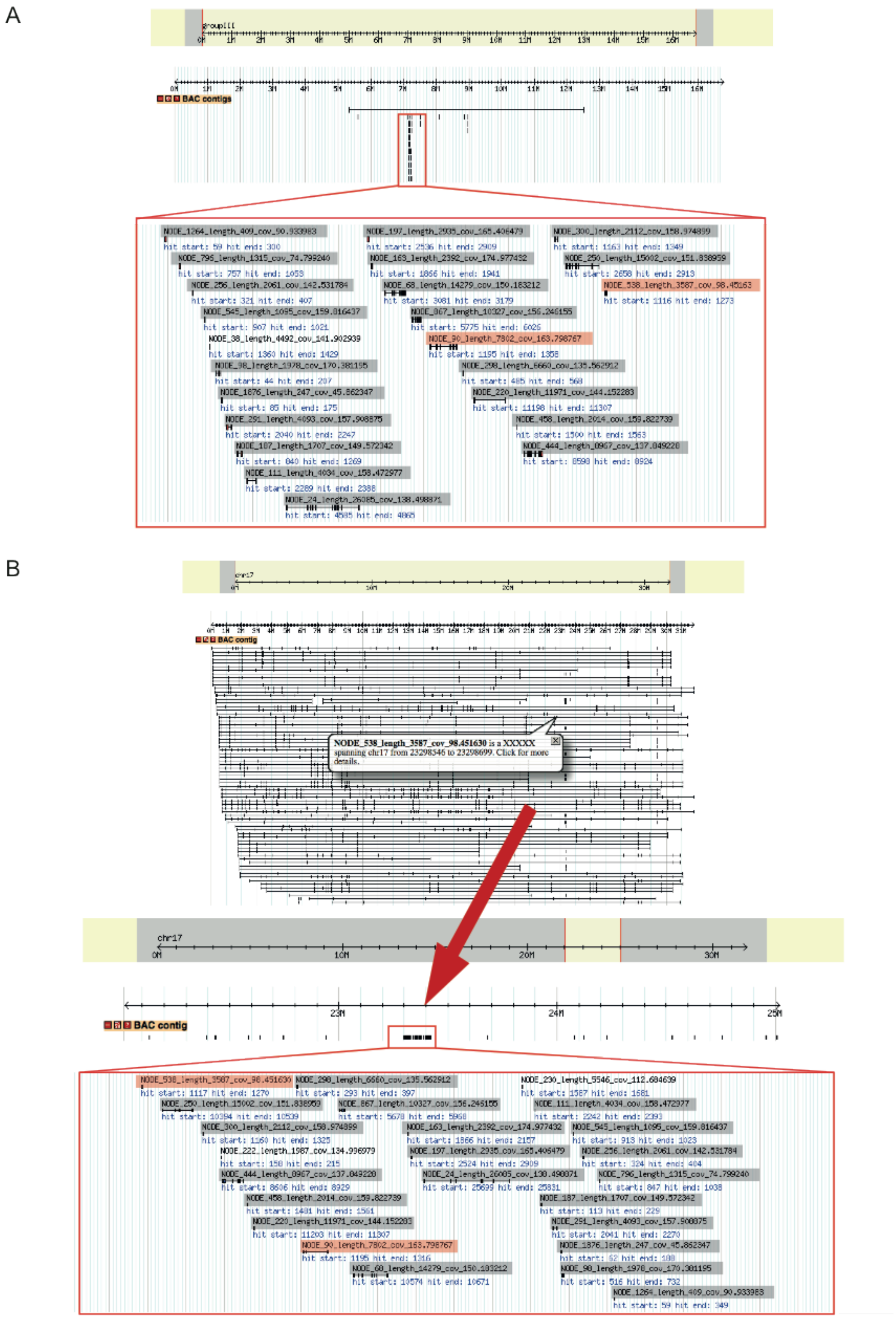

A
groupIII
BAC contigs
B
chr17
BAC contig
NODE_538_length_3587_cov_98.451630 is a XXXXX spanning chr17 from 23298545 to 23298699. Click for more details.
23M
24M
25M

Figure 5.1: Assigning contigs to BACs using GBrowse. All contigs were blasted against the stickleback and medaka reference genome and alignments were visualized in GBrowse. The search function was used in order to find the contigs containing Sanger sequenced parts of the BACs (see Supplementary Table S4.1). Contigs aligning densely to the same genomic region were believed to belong to the same BAC. This picture shows an example for Marker 0076. Contigs with significant hits on Sanger sequences (BAC ends, cDNA and genomic marker sequences) are marked in red. Contigs that have the same ordering on the reference genomes are marked in grey. A) Stickleback B) Medaka

Table 5.2 Summary of velvet assemblies using different hash values (*k*)

k	***exp_cov***	**Number of contigs**	**N50**	**Maximal contig length**	**Total output Mb**
31	692.3	469,567	27	21,221	6.18
33	663.5	4,316	173	1,318	0.64
35	634.6	3,543	192	2,049	0.59
37	605.8	3,621	203	1,783	0.63
39	576.9	4,073	221	2,190	0.77
41	548.1	4,293	230	1,802	0.85
43	519.2	4,615	241	2,087	0.95
45	490.4	82,833	904	29,312	3.29
47	461.5	65,245	1,801	34,866	3.09
49	432.7	50,668	3,042	32,728	2.9
51	403.8	31,472	4,945	29,668	2.61
53	375	20,124	6,819	38,458	2.39
55	346.2	13,751	5,241	24,728	2.25
57	317.3	10,871	2,505	18,186	2.13
59	288.5	7,797	2,552	16,632	2.07
61	259.6	5,761	2,729	22,623	2.02
63	230.8	4,477	2,717	22,621	1.99
65	201.9	4,483	2,717	22,621	1.99
67	173.1	4,500	2,717	22,621	1.99
69	144.2	4,502	2,717	22,621	1.99

Table 5.3 Summary of velvet assemblies using *k* = 53, but different values for the minimal and maximal coverage cutoffs.

t	**Number of contigs**	**N50**	**Maximal contig length**	**total output Mb**
5	845	4,963	18,971	1.23
10	1,504	6,511	34,592	1.74
15	1,738	8,184	38,468	1.93
20	2,096	8,557	38,468	1.95
25	2,515	8,484	38,659	1.97
30	2,778	8,589	38,661	1.97
35	3,136	8,589	38,657	1.98
40	3,654	8,589	38,657	2.00

the case of Marker 0290 and I will refer to this contig cluster as Cluster Bac04_G12/Bac33_A19. The total contig length in each cluster was between 95 and 194 Kb and in the range of the expected inserts size of the library. However, for Marker 0085 the total contig length was only 41 Kb (Table 5.4). The N50s and the maximal contig length for each contig cluster were very different and ranged from

2,842 to 34,644 bp and 4,001 to 38,520 bp, respectively (Table 5.4). In the following I will refer to each contig cluster by the BAC name (Table 5.4), e.g. Bac15_I02 refers to the contig cluster associated to Marker 0030.

Table 5.4: Statistics about contig clusters associated to BACs and the location of clusters in the two reference genomes of medaka and stickleback. Cov = average coverage, σ = standard deviation, LG = linkage group, Loc = location

Marker								**Medaka**		**Stickleback**	
No	**BAC name**	**number of contigs**	**average coverage**	**σ**	**N50**	**Max**	**total (Kb)**	**LG**	**Loc (Mb)**	**LG**	**Loc (Mb)**
30	**BAC15_I02**	11	108x	27	34,644	38,520	115.4	1	16.25	IX	1.24
76	**BAC25_P09**	24	140x	32	10,379	26,137	138.7	17	23.3	III	7.14
85	**BAC28_O05**	24	770x	190	2,842	4,011	41.2	8	23.3	XI	2.49
280	**BAC01O04**	20	58x	13	12,859	14,720	130	20	2.9	scaff 37	2.34
290	**BAC01_H09**	8	110x	11	10,531	30,569	84.2	7	0.58	XII	5.2
	/BAC32_G02	20	57x	12	4,431	10,571	88.6				
380	**BAC03_F10**	19	98x	20	17,434	34,236	142.4	21	20.03	XVI	12.9
396	**BAC03_I24**	22	146x	26	12,789	21,227	193.8	8	21.69	XI	2.53
455	**BAC04_G12**	21	318x	39	13,189	26,243	181.1	1	17.47	IX	9.69
/581	**/BAC33_A19**	-	-	-	-	-	-	-	-	-	-
1075	**BAC19_N24**	18	36x	5	10,464	16,464	95.5	12	8.15	XIV	3.44

5.2.3 Candidate genes

I found between three and eleven genes on the contigs in each cluster. With genes I refer to regions encoding well-annotated as well as hypothetical proteins. I found in total 66 genes of which 51 encode well-annotated protein. Well-annotated means here the function of the protein is known in other reference species. Relative gene order was mostly conserved across the two reference genomes of stickleback and medaka. However, some genes were in inverse order or were grouped together in one reference genome, but appeared in two separated genomic regions in the other reference genome. Separate means here the distance between genes was larger than 200 Kb and there were eventually other genes between them. In the first case, I could sometimes determine whether the respective region in the guppy is more similar to stickleback or medaka. In the latter case, I knew that the genes should be in the same region (here defined < 200 Kb) in the guppy genome, since they should be located on one BAC.

Marker 0030

BAC15_I02 contained in total ten genes, of which seven encode well-annotated proteins (Figure 5.2). I found all ten of them linked together in the same region in the stickleback reference genome. However, five of these genes were missing from the corresponding region in medaka, but the remaining genes were in the same order compared to stickleback. I was able to reconstruct the order for some genes in the guppy genome, since they were linked through contigs. From this I could infer that the ordering is mostly conserved between guppy and stickleback.

Marker 0076

I found nine genes on the contigs associated to Bac25_P09 and seven of them encode well-annotated proteins (Figure 5.2). Ordering of genes was conserved between the two reference species. Only one gene annotated as hypothetical protein in stickleback could not be found in the medaka reference, but had a hit on one of the guppy contigs. Those genes that could be ordered in the guppy appeared in the same relative order as compared to the stickleback and medaka.

Marker 0085

Here, I found seven genes in the respective region of medaka, of which five encode well-annotated proteins (Figure 5.2). In stickleback, I only found five of these genes, where three were found in another genomic region (further away than 150 Kb) on the same chromosome and the remaining two appeared to be in inverted order compared to medaka. The same five genes could also be found on the guppy contigs on Bac28_O05. The relative order of genes could not be determined, since none of them were linked through contigs. However, they should be linked together in the same genomic location similar to medaka.

Marker 0280

I found four genes on the contigs on Bac01_O04 of which three encode well-annotated proteins (Figure 5.2). One gene, *slc12A7*, could not be found in stickleback and another one, annotated as novel protein, could not be found in medaka. However, the order of the two remaining genes appeared to be conserved among the three species.

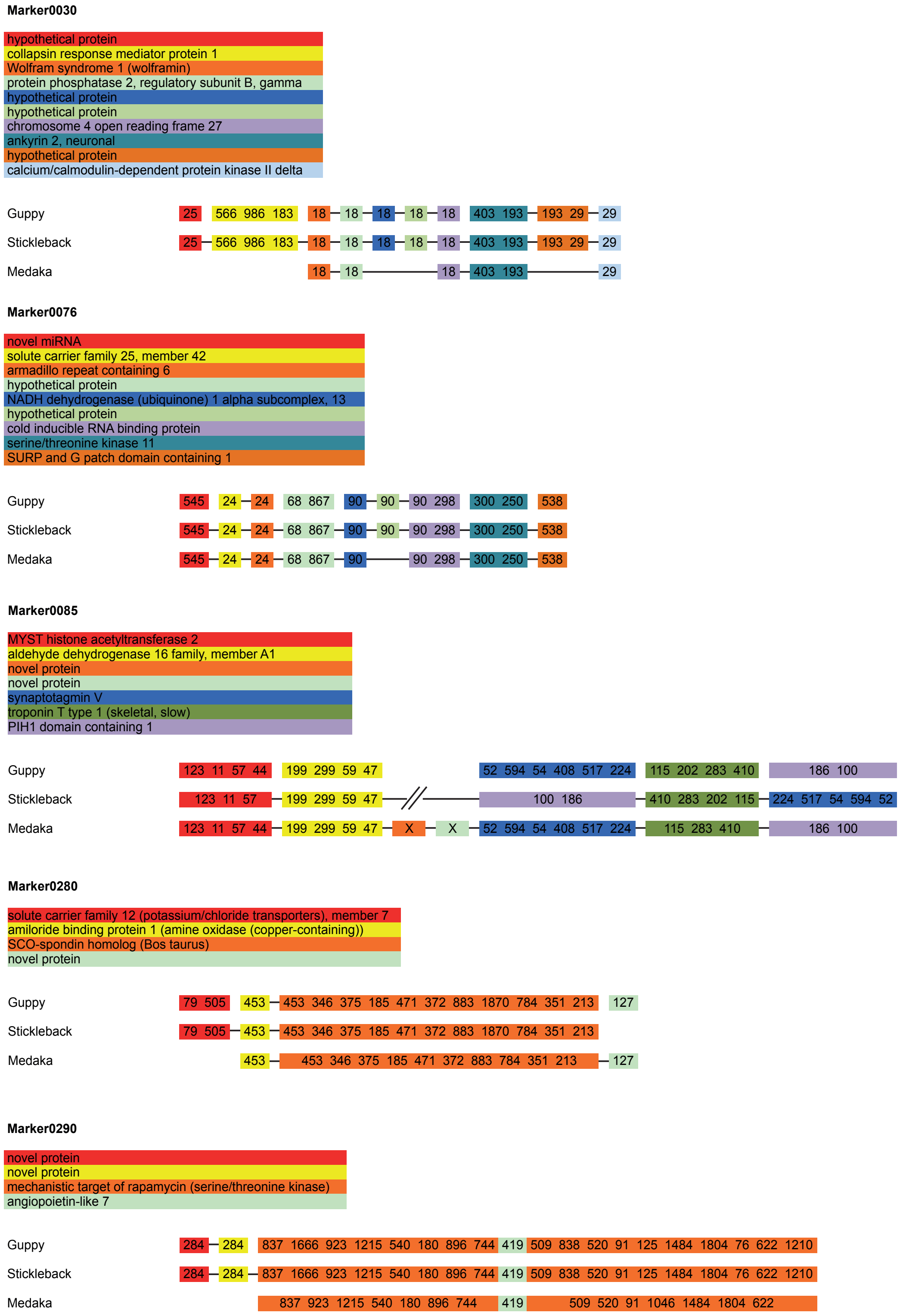


Figure 5.2: Comparative synteny analysis against two published fish genomes, medaka and stickleback. For each Marker all genes are listed that could be found in the contigs of the corresponding BAC. Genes are color coded and numbers within boxes refer to guppy contigs. In addition, the ordering of genes in the two reference genomes of stickleback and medaka is depicted using the same coloring as in lists as well as the most likely ordering in the guppy genome. If boxes are connected by a black line, the corresponding genes appear next to each other in the genome. Double slashes show that there is a gap between the sepeparated genes larger than 160 Kb.

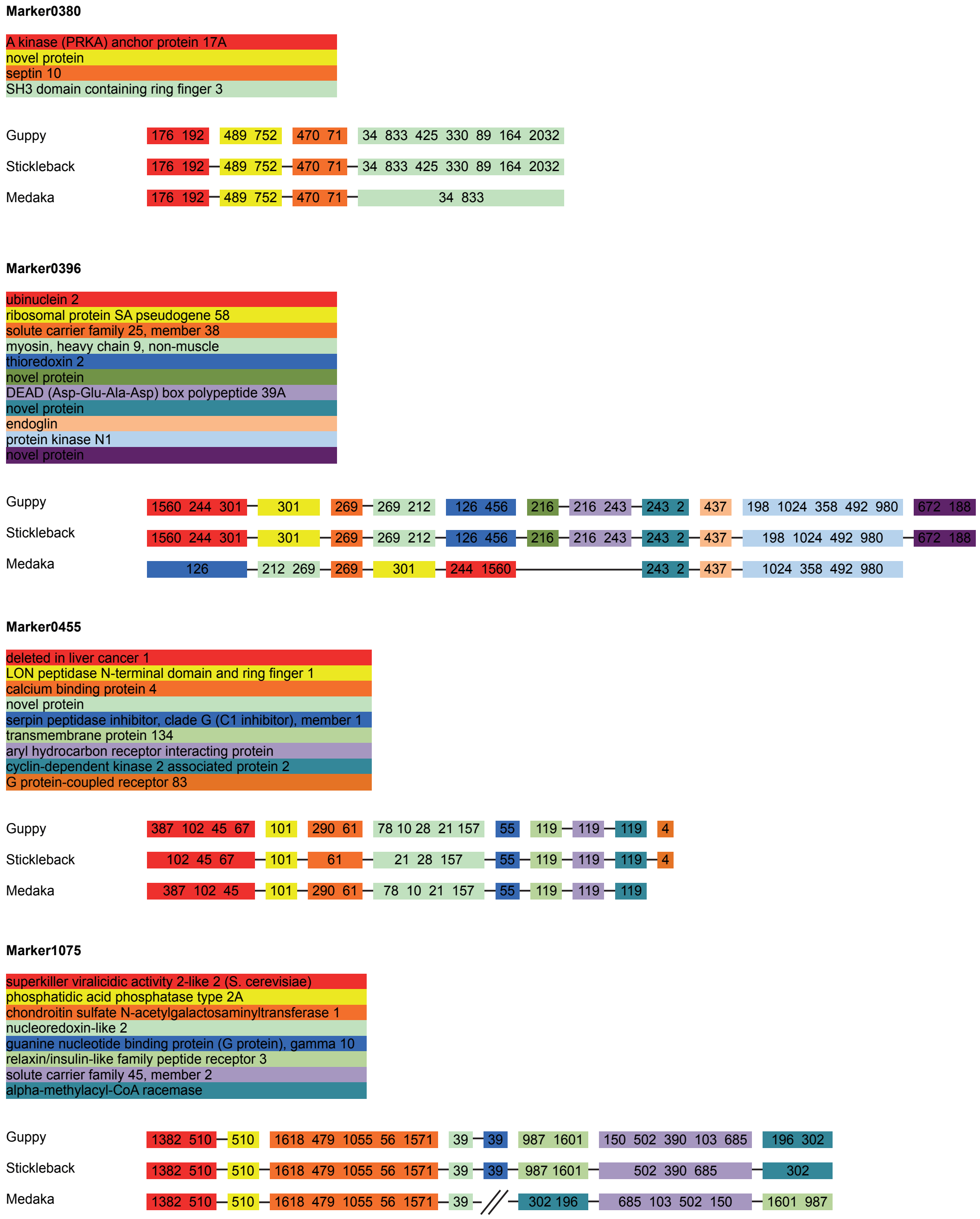


Figure 5.2: Comparative synteny analysis against two published fish genomes, medaka and stickleback. (continued)

Marker 0290

Marker 0290 was represented by two BACs, Bac01_H09 and Bac32_G02, but I combined the contigs into one cluster. These contigs contained four genes encoding two well-annotated and two hypothetical proteins (Figure 5.2). All four were linked together in the stickleback reference sequence, but two of them, both annotated as novel proteins, were missing in the reference sequence of medaka. Relative ordering of all four genes in the guppy genome could not be confidently resolved.

Marker 0380

I found four genes on the contigs associated to Bac03_F10,, of which only one is annotated as hypothetical protein. All four genes could be found in stickleback as well as medaka in conserved order. The relative ordering of these genes in the guppy genome could not be shown by contig links, but is most likely the same as in the two reference species.

Marker 0396

Bac03_I24 contained eleven genes and therefore the highest number of genes found on one BAC sequenced in this study. Eight encoded for well-annotated proteins (Figure 5.2). All of these genes could be found linked together in the stickleback reference sequence, but three genes were missing in medaka. In addition, a subset of the genes appeared to be inverted order in medaka relative to stickleback. I was not able to resolve whether the order of these genes in the guppy genome is more similar to medaka or stickleback. However, some of the genes could be linked through contigs and these subsets had the same relative order as in the two reference species.

Marker 0455/0581

As described before, the BACs associated to Marker 0455 and 0581 seem to be overlapping (Supplementary Figure S5.2). Therefore, contigs aligning to this region were fused into one cluster, Bac04_G12/Bac33_A19. The respective contigs contained nine genes of which eight encode well-annotated proteins (Figure 5.2). These genes could be found in conserved order in the stickleback and medaka reference sequence apart from the G protein coupled receptor 83 that was missing in medaka. Those genes, for which I was able to determine the order, appeared to have the same sequential arrangement in the guppy genome as in the medaka and stickleback reference sequence.

Marker 1075

The BAC associated to Marker 1075 contained an pigmentation candidate gene, namely the solute carrier family 45, member 2 (*slc45a2*). I found seven additional genes on Bac19_N24, all encoding well-annotated proteins (Figure 5.2). All of these could be found in the same genomic region of the stickleback reference, but in the medaka reference these genes were located in two different genomic locations. Moreover, one group appeared in reversed order compared to stickleback indicating an inversion (Figure 5.2). However, genes within this subgroup had the same relative order to each other compared to stickleback. Although, the order of genes in the guppy genome could not be confidently determined, it appeared that this genomic region is more similar to the syntenic region in the stickleback reference. The corresponding region containing all genes is too large in medaka to be covered by one BAC. This notion was supported by the observation that one of the contigs contained the breakpoint (Supplementary Figure S5.2).

5.3 Discussion

A pool of eleven BACs linked to ten genomic regions of interest has been sequenced without barcoding using Illumina GAII. *De-novo* assembly of the quality filtered dataset iterating over different parameters resulted in an assembly with an N50 of 8,557 bp and a maximum contig length of 38,468 bp. Previous studies using 454 technology reported N50s within the range of 3 to 50 Kb (Gonzalez, et al., 2010; Quinn, et al., 2008; Rounsley, et al., 2009; Sato, et al., 2011; Steuernagel, et al., 2009) using in some cases much longer read length and mate pair information. This indicates that our *de-novo* assembly using paired-end Illumina data was within the quality range of assemblies based on 454 data. However, the repeat content among different genomes significantly influences the quality and the N50 of *de-novo* assemblies and makes the results not directly comparable. Even within the genome the repeat content can vary drastically (e.g. near centromeres). Consequently, different parts of the genome might be easier to assemble and other parts are more difficult. This is also shown by our data. The repeat content among BACs seemed to vary significantly as shown by the different N50 values per contig cluster. Most contig cluster corresponding to BACs were found on the expected syntenic chromosomes in stickleback and medaka. Exceptions may have different explanations. If the probe used to screen the BAC library exists in multiple copies in the genome, it can happen that the wrong BAC is picked. For example, the BAC associated to Marker 0085 was probed with the cDNA of a gene existing in multiple copies in the medaka genome. Therefore, we might have picked a BAC containing a region that contains a paralogous gene. Another explanation might be that the syntenic genomic region is indeed located on a different chromosome in the guppy compared with stickleback and medaka. For example, Marker 0030 is located on guppy linkage group 12, which corresponds to the sex chromosome in guppy. However, I found that the contigs associated to Bac15_I02 through blast hits of the BAC ends do not align to the corresponding medaka chromosome 12, but mapped to medaka chromosome 1, which is the sex chromosome in this species. Yet, I would like to emphasize that the identification of BACs with their BAC ends is not unambiguous. Common repeats in the BAC ends can lead to wrong, but significant blast hits. Therefore, all results have to be interpreted with caution. This accounts especially for the BACs associated to Marker 0030, 0085 and 0380. Bac02_F10 associated to Marker 0380 could not even tagged by its BAC ends. It was just the only contig cluster left that has not been identified. It is unlikely that the wrong BAC was picked in this case, because it could be directly derived from the parent clone (Table 5.1). A BAC can loose large parts of its insert during DNA preparation due to repeats in the insert. This might be an explanation for a partly or completely missing BAC in the assembly.

Still, an important outcome of this study is that our approach can be successfully used for gene mining in a genomic region of interest, given that the correct BAC was picked. I found in total 66 genes linked to ten SNP markers of interest. Almost all genes, found in the syntenic region of medaka or stickleback, were covered by a contig in our assembly. However, the order of genes in the guppy could not always be fully resolved, because the assembly lacked sufficiently long continuous sequences. I did not use the scaffolding function in Velvet in order to generate longer scaffolds, because simulation studies have shown that scaffolding of pooled BAC data is very error prone (diploma thesis, Andrea Sprecher). Therefore, I preferred having less information about sequential arrangements of genes rather than getting wrong information in form of chimeric scaffolds. However, I cannot exclude that our assembly contains chimeric contigs in which parts actually come from different BACs. These chimeric contigs are most likely caused by repetitive motifs that are shared among different BAC sequences. The more BACs are pooled the higher the probability of shared repeat motifs among BACs. In this experiment, we did not barcode the BACs previously to pooling, in order to find out if this labor intensive step is necessary (Steuernagel, et al., 2009). Although I were able to assign 187 contigs to the different BACs, 529 contigs > 500 bp in the assembly remained of unknown origin. Contigs containing long stretches of coding sequence are more likely to be assigned correctly using synteny to other species. If the goal of a study using BAC sequencing is to find not only the coding parts of a BAC but also large parts of the non-coding stretches, I would recommend barcoding prior to pooling. With barcoding, reads could be sorted prior to assembly and each BAC could be assembled separately (Steuernagel, et al., 2009). In addition, it would provide the opportunity to optimize the assembly parameters for each BAC separately. This would most likely lead to better N50 values per BAC and would also solve the problem of chimeric contigs and scaffolds.

However, I was able to show that sequencing of BAC pools without barcoding using Illumina can be useful for identifying genes according to synteny to other species. Given that the BACs can be unambiguously identified, this approach can lead to interesting candidate genes linked to regions under selection. I found 44 genes on the seven BACs associated to SNP markers scored to be under directional selection (see Section 2.2.3). These genes could now be sequenced and analyzed further in different wild populations in order to determine whether or not some of them are indeed exposed to selection.

6. Conclusions

Studying genetic differences within and among natural populations can give important insights into ancestry, population structure, adaptation and speciation. Until recently genome-wide analyses were restricted to classical genetic model organisms, because for these the required molecular resources were available. However, there is not much known about the ecology of genetic model organisms and its difficult to infer from laboratory studies how natural populations adapt to their variable environment. Next generation sequencing makes it more feasible than ever to identify genetic loci responsible for adaptive evolution in ecological model organisms.

In this thesis I have shown by the means of the guppy, which is an important model organisms in ecological genetics, how new methods for high-throughput genotyping have shifted the field of population genetics to population genomics. I demonstrated the benefits of a comprehensive set of nuclear SNP markers anchored to the genetic map for a high-resolution survey of populations from a wide geographic range (Chapter 2). While I began with a comparatively low sample size per site, my study gave new insights in ancestry and population structure and specific questions can now be studied in depth using larger sample sizes from populations of interest. In addition I could show that, population parameters like F_{ST} were not inflated by small sample sizes (Chapter 3). As next-generation sequencing methods like RAD-seq became available, development of comprehensive SNP data sets are likely to become an efficient and cost-effective genetic tool in studies of non-reference species. I showed that my method especially designed for de-novo assembly and analysis of PE RAD-seq allowed to detect thousands of new genetic markers in the guppy genome suitable for genetic mapping and population studies (Chapter 4). RAD-seq will provide the opportunity to increase the density of informative markers in the guppy genome to the level required for resolution of genes under selection in contrasting, but geographically neighboring, habitats. Identification of genes under diversifying selection may guide the choices of genomic regions to be sequenced with BAC-seq (Chapter 5) and help us decipher the evolution of adaptation in natural guppy populations.

During the time of this thesis, the field of population genetics has shifted to next generation population genomics. We started out in the guppy with a set of 1,000 SNP markers, which was a huge number of genetic markers for a non-reference organism four years ago. Meanwhile, we are able to generate thousands of new SNP markers at only a fraction of cost using NGS. As mentioned in Chapter 2, analyses of previous SNP datasets were suffering from ascertainment bias (Luikart, et al., 2003), , which can significantly bias the estimate of F_{ST} (Chapter 3). This problem is solved with next generation genotyping methods like RAD-seq, because all individuals included in the study can be genotyped at the same time.

Interesting questions in future population genetic studies in non-reference organisms will no longer depend on the number of available markers, because population genomics will become feasible in a large variety of systems. However, the experimental design of these studies depends on the question of interest. In Chapter 3, I showed that taking a very small number of individuals from each population led already to reliable estimates of genetic differentiation. Therefore, a study on population substructure and demographic histories might be done now with a large number of populations from a wide geographic range at very little cost.

However, explaining the mechanisms of adaptation and speciation is one of the major goals in evolutionary genetics. In these studies large sample sizes from the populations analysed are required, because allele frequencies at single loci have to be estimated (Chapter 3). Still, we now have the tools to analyse adaptive divergence genetically in a large number of different systems hopefully giving us answers to questions dating back to the early days of evolutionary biology. There will be an explosion of new studies aiming at identifying genes responsible for adaptive change, because empirical approaches for F_{ST} outlier methods become feasible in non-model taxa. Studies similar to the one done by Akey et al. (2002), who used > 26,000 SNPs sampled genome wide in humans to find regions under selection, will now be possible in non-reference taxa (Luikart, et al., 2003).

The vast amount of genotype data that will soon be available will also bring new challenges for the field of Bioinformatics. Many of the commonly used software programs for population genetic data analysis (e.g. (Excoffier and Heckel, 2006) cannot deal with large SNP datasets. The computation times of the sophisticated software STRUCTURE (see Chapter 2) scales with the number of individuals and markers used as input. With thousands of markers, it will not be feasible to run STRUCTURE. PCA (see Chapter 2) is a valid alternative to STRUCTURE in order to infer population substructure. The computation times are very fast even if using thousands of markers. Moreover, it provides the opportunity to sort the SNPs according to information content (Paschou, et al., 2007). This feature could be used to choose a subset of SNPs used for STRUCTURE in order to infer admixture patterns. This is only one example of how already available software packages can be combined in order to be able to deal with the huge datasets. However, also new software adapting old methods to the new data is needed, in addition to completely new approaches developed to take full advantage of the provided genome-wide information. Concisely, software development in this field has to move from population genetics to next generation population genomics.

However, next generation population genomics will revolutionize the field of evolutionary genetics and I am positive that my work will be helpful in providing some guidelines and new methods for future studies.

References

Achere, V.*, et al.* (2005) Genomic organization of molecular differentiation in Norway spruce (*Picea abies*), *Molecular Ecology*, **14**, 3191-3201.

Akey, J.M.*, et al.* (2002) Interrogating a high-density SNP map for signatures of natural selection, *Genome Research*, **12**, 1805-1814.

Alexander, H. and Breden, F. (2004) Sexual isolation and extreme morphological divergence in the Cumana guppy: a possible case of incipient speciation., *Journal of Evolutionary Biology*, **17**, 1238 - 1254.

Alexander, H.*, et al.* (2006) Parallel evolution and vicariance in the guppy (*Poecilia reticulata*) over multiple spatial and temporal scales, *Evolution*, **60**, 2353 - 2369.

Altschul, S.F.*, et al.* (1990) Basic local alignment search tool, *Journal of Molecular Biology*, **215**, 403-410.

Baird, N.A.*, et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers, *PLoS One*, **3**, e3376.

Barson, N.J., Cable, J. and Van Oosterhout, C. (2009) Population genetic analysis of microsatellite variation of guppies (*Poecilia reticulata*) in Trinidad and Tobago: evidence for a dynamic source-sink metapopulation structure, founder events and population bottlenecks, *Journal of Evolutionary Biology*, **22**, 485-497.

Beaumont, M.A. and Balding, D.J. (2004) Identifying adaptive genetic divergence among populations from genome scans, *Molecular Ecology*, **13**, 969-980.

Beaumont, M.A. and Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure, *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **363**, 1619-1626.

Becher, S.A. and Magurran, A.E. (1999) Gene flow in Trinidadian guppies, *Journal of Fish Biology*, **56**, 241 - 249.

Bensch, S. and Akesson, M. (2005) Ten years of AFLP in ecology and evolution: why so few animals?, *Molecular Ecology*, **14**, 2899-2914.

Bonin, A.*, et al.* (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*), *Molecular Biology and Evolution*, **23**, 773-783.

Breden, F., Scott, M. and Michel, E. (1987) Genetic differentiation for antipredator behaviour in the Trinidadian guppy, *Poecilia reticulata*, *Animal Behaviour*, **35**, 618 - 620.

Briggs, J.C. (1984) Freshwater fishes and biogeography of Central America and the Antilles, *Systematic Zoology*, **33**, 428 - 435.

Bryant, D. and Moulton, V. (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks, *Molecular Biology and Evolution*, **21**, 255-265.

Campbell, D. and Bernatchez, L. (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes, *Molecular Biology and Evolution*, **21**, 945-956.

Carvalho, G.R.*, et al.* (1991) Marked genetic divergence revealed by allozymes among populations of the guppy *Poecilia reticulata* (Poecilidae), in Trinidad, *Biological Journal of the Linnean Society*, **42**, 389 - 405.

Chaisson, M.J. and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes, *Genome Research*, **18**, 324-330.

Crispo, E.*, et al.* (2006) The relative influence of natural selection and geography on gene flow in guppies, *Molecular Ecology*, **15**, 49-62.

Di Rienzo, A.*, et al.* (1994) Mutational processes of simple-sequence repeat loci in human populations, *Proceedings of the National Academy of Sciences*, **91**, 3166-3170.

Dreyer, C.*, et al.* (2007) ESTs and EST-linked polymorphisms for genetic mapping and phylogenetic reconstruction in the guppy, Poecilia reticulata, *BMC Genomics*, **8**, 269.

Elshire, R.J.*, et al.* (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species, *PLoS One*, **6**, e19379.

Endler, J.A. (1978) A predator's view of animal color patterns, *Evolutionary Biology*, **11**, 319 - 364.

Endler, J.A. (1980) Natural selection on color patterns in *Poecilia reticulata*, *Evolution*, **34**, 76 - 91.

Endler, J.A. (1983) Natural and sexual selection on color patterns in poeciliid fishes, *Enviromental Biology of Fishes*, **9**, 173 - 190.

Endler, J.A. (1991) Variation in the appearance of guppy color patterns to guppies and their predators under different visual conditions, *Vision Research*, **31**, 587-608.

Endler, J.A. (1995) Multiple-Trait coevolution and environmental gradients in guppies *Trends in Ecology and Evolution*, **10**, 22 - 29.

Estoup, A.*, et al.* (1995) Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae), *Molecular Biology and Evolution*, **12**, 1074-1084.

Excoffier, L. (2007) Analysis of Population Subdivision. In Balding, D.J., Bishop, M. and Cannings, C. (eds), *Handbook of Statistical Genetics*. John Wiley & Sons, Ltd., West Sussex, pp. 982pp.

Excoffier, L. and Heckel, G. (2006) Computer programs for population genetics data analysis: a survival guide, *Nature Reviews Genetics*, **7**, 745-758.

Excoffier, L., Laval, G. and Schneider, S. (2005) arlequin version 3.0: an integrated software package for population genetics data analysis, *Evolutionary Bioinformatics Online*, **1**, 47 - 50.

Fajen, A. and Breden, F. (1992) Mitochondrial-DNA sequence variation among natural populations of the Trinidad guppy, *Poecilia reticulata*, *Evolution*, **46**, 1457 - 1465.

Falush, D., Stephens, M. and Pritchard, J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, *Genetics*, **164**, 1567-1587.

Ghalambor, C.K., Reznick, D.N. and Walker, J.A. (2004) Constraints on adaptive evolution: the functional trade-off between reproduction and fast-start swimming performance in the Trinidadian guppy (*Poecilia reticulata*), *American Naturalist*, **164**, 38 - 50.

Gnerre, S*., et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data, *Proceedings of the National Academy of Sciences U S A*, **108**, 1513-1518.

Godin, J.-G. and McDonough, H.E. (2003) Predator preference for brightly colored males in the guppy: a viability cost for a sexually selected trait, *Behavioral Ecology* **14**, 194 - 200.

Gonzalez, V.M*., et al.* (2010) Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy, *BMC Plant Biology*, **10**, 246.

Hartigan, J.A. and Wong, M.A. (1979) A K-means clustering algorithm, *Applied Statistics*, **28**, 128 - 137.

Hartl, D.L. and Clark, A.G. (1997) *Principles of Population Genetics*. Sinauer Associates.

Hedrick, P.W. (1999) Perspective : Highly variable loci and their interpretation in evolution and conservation, *Evolution*, **53**, 313-318.

Hoffmann, J.I. and Amos, W. (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion, *Molecular Ecology*, **14**, 599-612.

Hohenlohe, P.A*., et al.* (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout, *Molecular Ecology Resources*, **11 Suppl 1**, 117-122.

Holsinger, K.E. and Weir, B.S. (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST), *Nature Reviews Genetics*, **10**, 639-650.

Houde, A.E. (1997) Sex, Color and Mate Choice in Guppies., *Princeton University Press, Princeton, New Jersey*.

Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution*, **23**, 254-267.

Jakobsson, M. and Rosenberg, N.A. (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, *Bioinformatics (Oxford, England)*, **23**, 1801-1806.

Karim, N., *et al.* (2007) This is not deja vu all over again: male guppy colour in a new experimental introduction., *Journal of Evolutionary Biology*, **20**, 1339-1350.

Kelley, J.L. and Magurran, A.E. (2003) Effects of relaxed predation pressure on visual predator recognition in the guppy, *Behavioral Ecology and Sociobiology*, **54**, 225 - 232.

Kenny, J.S. (1988) Hermatypic scleractinian corals of Trinidad., *Studies of the fauna of Curacao and other Caribbean islands*, **123**, 83-100.

Klein, J.D., *et al.* (2011) LOCAS--a low coverage assembly tool for resequencing projects, *PLoS One*, **6**, e23455.

Lewontin, R.C. and Krakauer, J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms, *Genetics*, **74**, 175-195.

Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome Research*, **18**, 1851-1858.

Luikart, G., *et al.* (2003) The power and promise of population genomics: from genotyping to genome typing, *Nature Reviews Genetics*, **4**, 981-994.

Luikart, G., *et al.* (2003) The power and promise of population genomics: from genotyping to genome typing, *Nature Reviews Genetics*, **4**, 981-994.

Magurran, A.E. (1998) Population differentiation without speciation, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **353**, 275 - 286.

Magurran, A.E. (2005) *Evolutionary Ecology: The Trinidadian Guppy*. Oxford University Press, Oxford.

Magurran, A.E. and Seghers, B.H. (1990) Population differences in predator recognition and attack cone avoidance in the guppy *Poecilia reticulata*, *Animal Behaviour*, **40**, 443 - 452.

Magurran, A.E*., et al.* (1992) Behavioural consequences of an artificial introduction og guppies (*Poecilia reticulata*) in N. Trinidad: evidence for the evolution of ant-predator behaviour in the wild, *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **248**, 117 - 122.

Magurran, A.E*., et al.* (1995) The behavioural diversity and the evolution of guppy, *Poecilia reticulata*, populations in Trinidad, *Advances in the Study of Behaviour*, **24**, 155 - 202.

Mäkinen, H.S., Cano, J.M. and Merila, J. (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations, *Molecular Ecology*, **17**, 3565-3582.

Malhis, N. and Jones, S.J. (2010) High quality SNP calling using Illumina data at shallow coverage, *Bioinformatics*, **26**, 1029-1035.

Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics, *Trends in Genetics*, **24**, 133-141.

Margulies, M*., et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376-380.

Martin, C.H. and Johnson, S. (2007) A field test of the Hamilton-Zuk hypothesis in the Trinidadian guppy (*Poecilia reticulata*), *Behavioral Ecology and Sociobiology*, **61**, 1897-1909.

Miller, C.R., Joyce, P. and Waits, L.P. (2002) Assessing allelic dropout and genotype reliability using maximum likelihood, *Genetics*, **160**, 357-366.

Morin, P.A*., et al.* (2004) SNPs in ecology, evolution and conservation, *Trends in Ecology and Evolution*, **19**, 208-216.

Morin, P.A., Martien, K.K. and Taylor, B.L. (2009) Assessing statistical power of SNPs for population structure and conservation studies, *Molecular Ecology Resources*, **9**, 66-73.

Namroud, M.C*., et al.* (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce, *Molecular Ecology*, **17**, 3599-3613.

Narum, S.R*., et al.* (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms, *Molecular Ecology*, **17**, 3464-3477.

Nordborg, M*., et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*, *PLoS Biology*, **3**, e196.

O'Steen, S., Cullum, A.J. and Bennett, A.F. (2002) Rapid evolution of escape ability in Trinidadian guppies (*Poecilia reticulata*), *Evolution*, **56**, 776 - 784.

Paschou, P*., et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations, *PLoS Genetics*, **3**, 1672-1686.

Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis, *PLoS Genetics*, **2**, e190.

Pfender, W.F*., et al.* (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in Lolium perenne, *Theoretical Applied Genetics*.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945-959.

Pritchard, J.K. and Wen, W. (2002) Documentation for structure software: Version 2. *Department of Human Genetics*. University of Chicago, Chicago.

Quinn, N.L*., et al.* (2008) Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome, *BMC Genomics*, **9**, 404.

Reznick, D.N. and Bryga, H.A. (1987) Life-history evolution in guppies (*Poecilia reticulata*). 1. Phenotypic and genetic changes in an introduction experiment, *Evolution*, **41**, 1370 - 1385.

Reznick, D.N. and Bryga, H.A. (1996) Life-history evolution in guppies (*Poecilia reticulata*: Poeciliidae). V. Genetic basis of parallelism in life histories, *American Naturalist*, **147**, 339-359.

Reznick, D.N., Rodd, F.H. and Cardenas, M. (1990) Experimentally induced life-history evolution in a natural population, *Nature*, **346**, 357 - 359.

Reznick, D.N., Rodd, F.H. and Cardenas, M. (1996b) Life-history evolution in guppies (*Poecilia reticulata*: Poecilidae). IV. Parallelism in life-history phenotypes, *American Naturalist*, **147**, 319 - 338.

Rosenblum, E.B. and Novembre, J. (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard, *Journal of Heredity*, **98**, 331-336.

Rounsley, S.*, et al.* (2009) De novo next generation sequencing of plant genomes, *Rice*, **2**, 35-43.

Salmela, E.*, et al.* (2008) Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe, *PLoS One*, **3**, e3519.

Sato, K.*, et al.* (2011) 454 sequencing of pooled BAC clones on chromosome 3H of barley, *BMC Genomics*, **12**, 246.

Schneeberger, K.*, et al.* (2009) Simultaneous alignment of short reads against multiple genomes, *Genome Biology*, **10**, R98.

Seddon, J.M.*, et al.* (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population, *Molecular Ecology*, **14**, 503-511.

Seghers, B.H. (1974) Schooling behavior in the guppy (*Poecilia reticulata*): an evolutionary response to predation, *Evolution*, **28**, 486 - 489.

Shaw, P.W.*, et al.* (1991) Population differentiation in Trinidadian guppies (*Poecilia reticulata*): patterns and problems, *Journal of Fish Biology*, **39**, 203 - 209.

Shaw, P.W.*, et al.* (1992) Genetic Consequences of an artificial introduction of Guppies (*Poecilia reticulata*) in N. Trinidad, *Proceedings: Biological Science*, **248**, 111 - 116.

Stein, L.D.*, et al.* (2002) The generic genome browser: a building block for a model organism system database, *Genome Research*, **12**, 1599-1610.

Steuernagel, B.*, et al.* (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley, *BMC Genomics*, **10**, 547.

Suk, H.Y. and Neff, B.D. (2009) Microsatellite genetic differentiation among populations of the Trinidadian guppy, *Heredity*, **102**, 425-434.

Summerer, D. (2009) Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing, *Genomics*, **94**, 363-368.

Tautz, D., Ellegren, H. and Weigel, D. (2010) Next generation molecular ecology, *Molecular Ecology*, **19 Suppl 1**, 1-3.

Team, R.D.C. (2008) R: A Language and Enviroment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Tripathi, N.*, et al.* (2009) Genetic linkage map of the guppy, *Poecilia reticulata*, and quantitative trait loci analysis of male size and colour variation, *Proceedings. Biological sciences / The Royal Society*, **276**, 2195-2208.

van Oosterhout, C., Harris, P.D. and Cable, J. (2003) Marked variation in parasite resistance between two wild populations of the Trinidadian guppy, *Poecilia reticulata* (Pisces: Poeciliidae), *Biological Journal of the Linnean Society*, **79**, 645-651.

van Oosterhout, C.*, et al.* (2007) The guppy as a conservation model: implications of parasitism and inbreeding for reintroduction success, *Conservation Biology*, **21**, 1573-1583.

van't Hof, A.E.*, et al.* (2011) Industrial melanism in British peppered moths has a singular and recent mutational origin, *Science*, **332**, 958-960.

Vasemagi, A., Nilsson, J. and Primmer, C.R. (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in atlantic salmon (*Salmo salar* L.), *Molecular Biology and Evolution*, **22**, 1067-1076.

Vignal, A.*, et al.* (2002) A review on SNP and other types of molecular markers and their use in animal genetics, *Genetics, Selection, Evolution: GSE*, **34**, 275-305.

Weir, B.S. (1996) *Genetic Data Analysis II*. Sinauer Associates Inc., Sunderland, MA.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure, *Evolution*, **38**, 1358-1370.

Wicker, T.*, et al.* (2006) 454 sequencing put to the test using the complex genome of barley, *BMC Genomics*, **7**, 275.

Willing, E.M.*, et al.* (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies, *Molecular Ecology*, **19**, 968-984.

Willing, E.M.*, et al.* (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference, *Bioinformatics*, **27**, 2187-2193.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Research*, **18**, 821-829.

Contributions

Chapter 1: Introduction

Some parts from the Introduction are taken from Willing et al. (2010), Willing et al. (2011) and Willing et al. (*in preparation*).

Chapter 2: Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies

This work was published in Molecular Ecology (Willing, et al., 2010) with the following authors Eva-Maria Willing (EMW), Paul Bentzen (PB), Cock van Oosterhout (CvO), Margarete Hoffmann (MH), Joanne Cable (JC), Felix Breden (FB), Detlef Weigel (DW) and Christine Dreyer (CD). CD and DW conceived the study. EMW designed the study. MH did all the wet lab work that was needed to genotype the populations analyzed. PB, CvO, JC and FB provided samples of wild populations analyzed. EMW chose all the methods used and performed all analyses leading to the results described. EMW implemented new methods where needed using Perl and Java. CD and EMW wrote the manuscript, which was critically revised and approved by all authors.

Permissions and Copyright agreement Molecular Ecology: Authors are allowed to reuse the content of their own articles.

Chapter 3: Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using large panel of SNP markers

This work is under preparation for publication with the following authors, Eva-Maria Willing (EMW), Christine Dreyer (CD) and Cock van Oosterhout (CvO). EMW, CD and CvO conceived the study. EMW designed the study. EMW developed the software needed for the simulation and performed the research. EMW wrote the manuscript that was critically revised by CD and CVO.

Chapter 4: Paired-end RAD-seq for de-novo assembly and marker design without available reference

This work has been published in Bioinformatics (Willing, et al., 2011) with the following authors: Eva-Maria Willing (EMW), Margarete Hoffmann (MH), Juliane D. Klein (JK), Detlef Weigel (DW), Christine Dreyer (CD). CD and EMW conceived the study. EMW, MH and CD designed the experiment. MH did the wet lab work. EMW designed and JK implemented LOCASopt. EMW developed the remaining software written in Perl and available as package RApiD. EMW performed the research. EMW wrote the manuscript, which was critically revised by CD and DW and approved by all authors.

Permissions and Copyright agreement Bioinformatics: It is authors permitted to include the article in full or in part in a thesis or dissertation, provided that this not published commercially.

Chapter 5: Sequencing of specific genomic regions contained in BACs with short reads

This work has not been published. Eva-Maria Willing (EMW), Christine Dreyer (CD), Margarete Hoffmann (MH) and Verena Kottler (VK) participated in this study. EMW conceived the study. EMW and CD designed the experiment. EMW, CD and VK selected the BACs for sequencing. MH and VK performed the wet lab work. EMW did the all analyses of the sequencing data. EMW wrote the chapter.

Curriculum Vitae

Nachname:	**Willing**
Vorname:	**Eva-Maria**
Geburtsdatum:	**12.09.1979**
Geburtsort:	**Vreden**
Nationalität:	**Deutsch**

AUSBILDUNG

05/2011 – jetzt	**Max Planck Institut für Planzenzüchtungsforschung** • **Forschungsgruppe:** Genome Plasticity and Computational Genetics
03/2007 – 04/2011	**Max Planck Institut für Entwicklungsbiologie** • **Forschungsgruppe:** Variation and adaptation in the guppy, *Poecilia reticulata* (www.weigelworld.org/research/projects/guppyvariation) • **Betreuer:** Christine Dreyer, Detlef Weigel, Daniel Huson
10/2000 – 02/2007	**Eberhard-Karls University of Tübingen, Germany** • **Abschluss**: Diplom Informatik (Bioinformatik) (Note: sehr gut) • **Diplomarbeit:** “Molecular Phylogeny of Guppy” at the Max Planck Institut für Entwicklungsbiologie und Universität Tübingen • **Betreuer:** Christine Dreyer, Kay Nieselt, Stephan R. Henz
08/1990 – 06/1999	**Gymnasium Georgianum in Vreden** **Abschluss**: Abitur (Note: gut)

PUBLIKATIONEN

Dreyer C, Hoffmann M, Lanz C, **Willing EM**, Riester M, Warthmann N, Sprecher A, Tripathi N, Henz SR, Weigel D (2007) ESTs and EST-linked polymorphisms for genetic mapping and phylogenetic reconstruction in the guppy, *Poecilia reticulata*, *BMC Genomics*, 8, 269

Tripathi N, Hoffmann M, **Willing EM**, Lanz C, Weigel D, Dreyer C. (2009) Genetic linkage map of the guppy, *Poecilia reticulata*, and quantitative trait loci analysis of male size and colour variation, *Proceedings. Biological Sciences*, 276, 2195-2208.

Willing EM, Bentzen P, van Oosterhout C, Hoffmann M, Cable J, Breden F, Weigel D, Dreyer C. (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies, *Molecular Ecology*, 19, 968-984

Willing EM, Hoffmann M, Klein JD, Weigel D, Dreyer C. (2011a) Paired-end RAD-seq for *de-novo* assembly and marker design without available reference, *Bioinformatics*, **27**, 2187-2193

Willing EM, Dreyer C, van Oosterhout C. (*in preparation*) Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes if using thousands of bi-allelic markers

VORTRÄGE

STEvE (Tübingen, 2007), Title: Analysing the molecular phylogeny of the Guppy (*Poecilia reticulata*) using modern phylogenetic methods

3rd European Poeciliid Conference (Chioggia, 2008), Title: Population structure in wild guppy populations

STEvE (Tübingen, 2008), Title: Analysis of high throughput SNP genotyping reveals complex hierarchal sub-structure in guppies (*Poecilia reticulata*)

PhD Day (MPI, Tübingen, 2009), Title: Adaptation of wild guppies - Molecular Analyses

Summer School "Ecological Genomics" (Bertinoro, 2009), Title: Adaptation of wild guppies - Molecular Analyses

4th European Poeciliid Conference (St. Andrews, 2010), Title: Looking for regions under selection in the guppy genome

TIPP Retreat (Schwäbische Alb, 2010), Title: Looking for regions under selection in the guppy genome (Award for best talk)

Symposium: "Evolutionary and ecological genomics of adaptation" (Fribourg, 2010) "Looking for regions under selection in the guppy genome"

Fish Genome Meeting (Hinxton, 2011) "Finding new SNP markers in the guppy genome by analyzing RAD tags without using a reference"

DiKo (MPI, Tübingen, 2011), Title: Finding regions under selection in the guppy genome

Appendix

A. Supplementary Figures

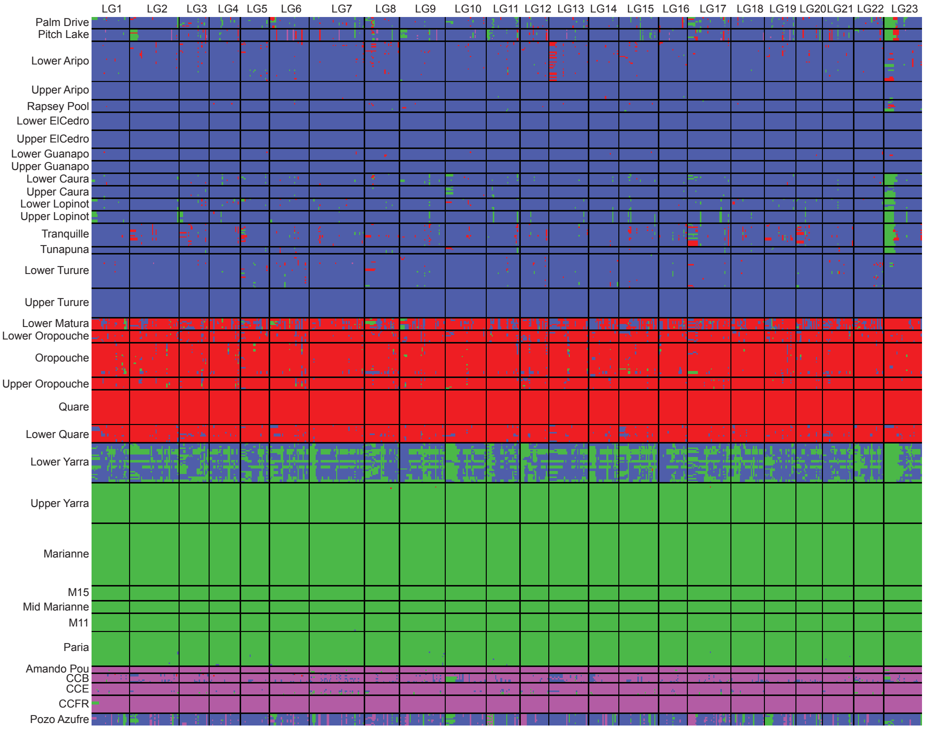


Figure S2.8: Linkage plot. Clustering was inferred with STRUCTURE using the linkage model with 100,000 burn in steps and 100,000 iteration steps with k = 4. Each individual is represented by two subsequent rows and columns represent genotypes ordered according to the position on the linkage groups. Colors describe the most likely cluster membership for each genotype.

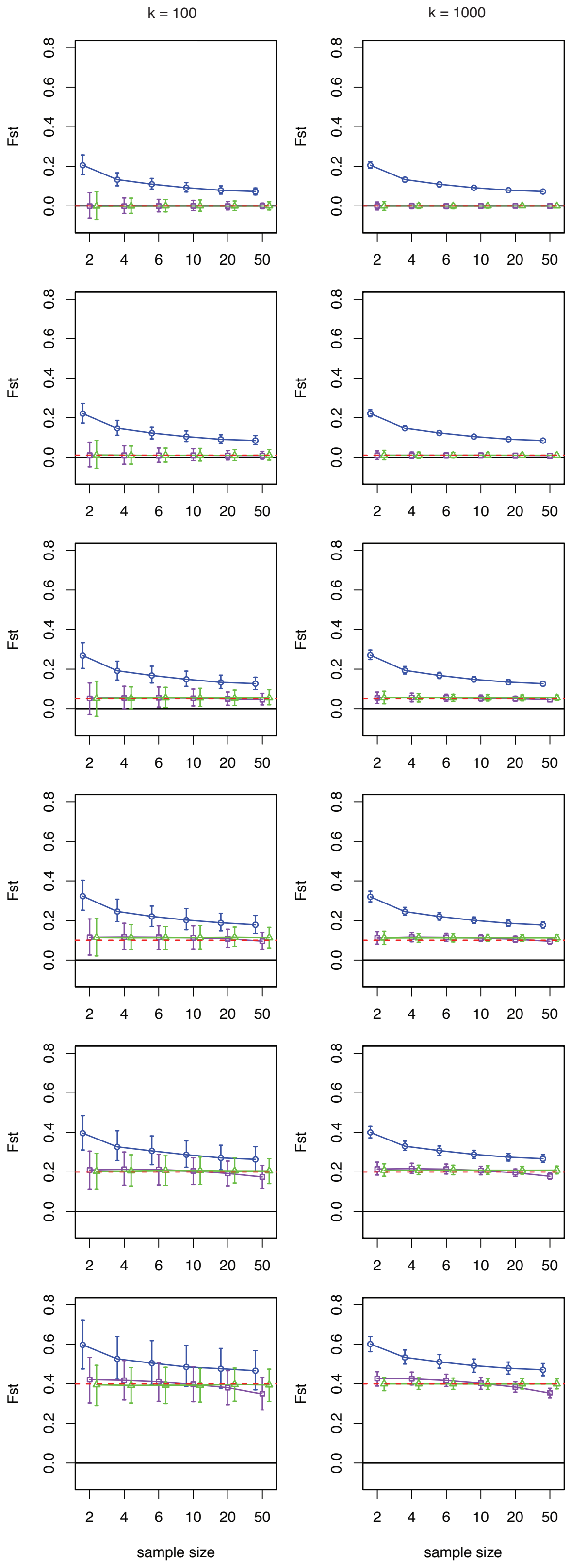


Figure S3.1: Effect of increasing sample sizes. Results are shown for the simulated data with equally distributed allele frequencies using different sample sizes taken from population 1 (n_1 = 4) and from population 2 (n_2 = 2, 4, 6, 10, 20, 50). Number of loci is fixed at k = 100 (left column) and k = 1,000 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: F_{ST}^{W} (blue circles), $F_{ST}^{W\&C}$ (purple squares) and F_{ST}^{R} (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

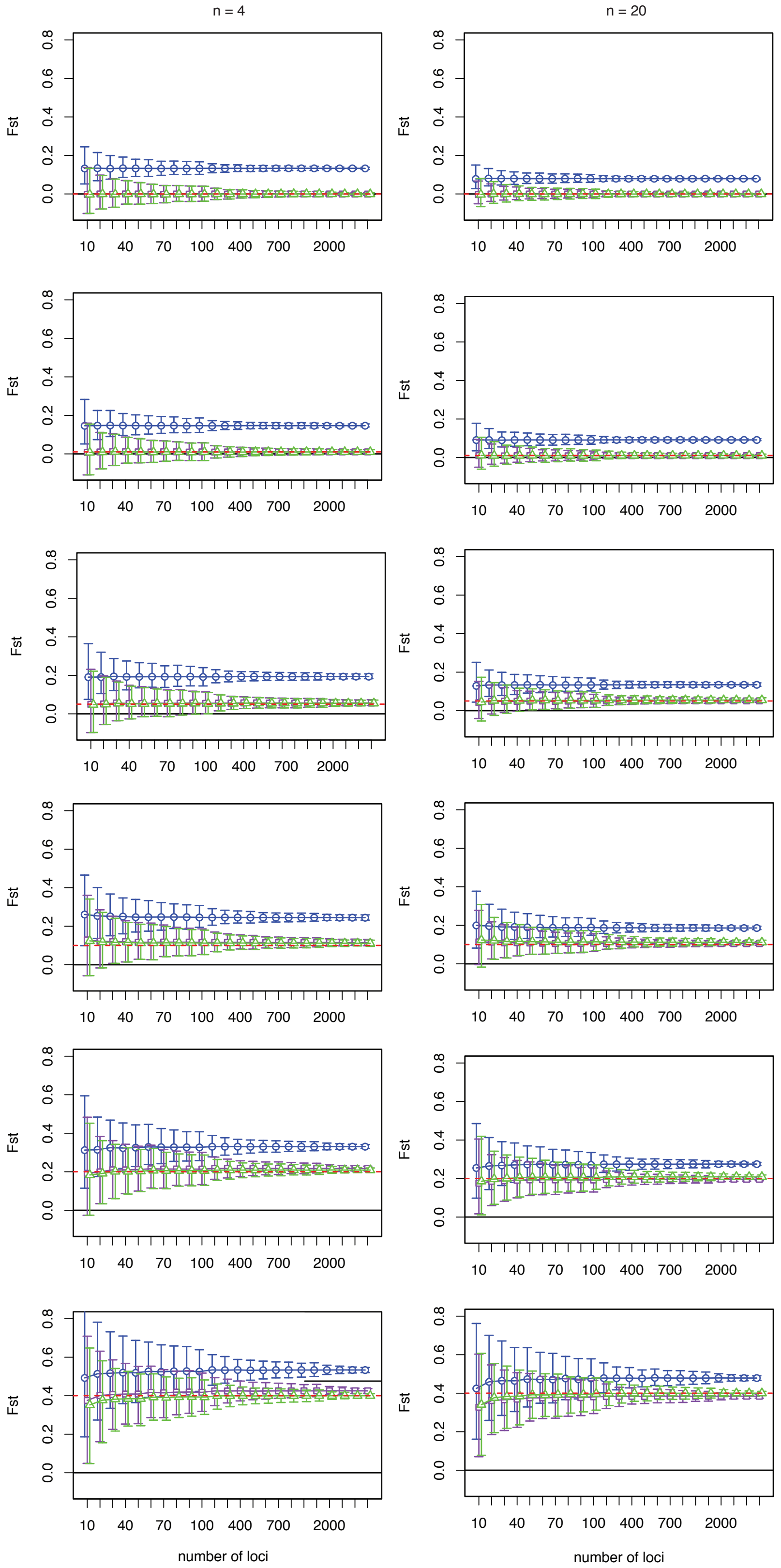


Figure S3.2: Effect of increasing the number of markers. Results are shown for the simulated data with equally distributed allele frequencies using different sample sizes taken from population 1 ($n_1 = 4$) and from population 2 ($n_2 = 2, 4, 6, 10, 20, 50$). Number of individuals is fixed at n = 4 (left column) and n = 20 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: F_{ST}^{W} (blue circles), $F_{ST}^{W\&C}$ (purple squares) and F_{ST}^{R} (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

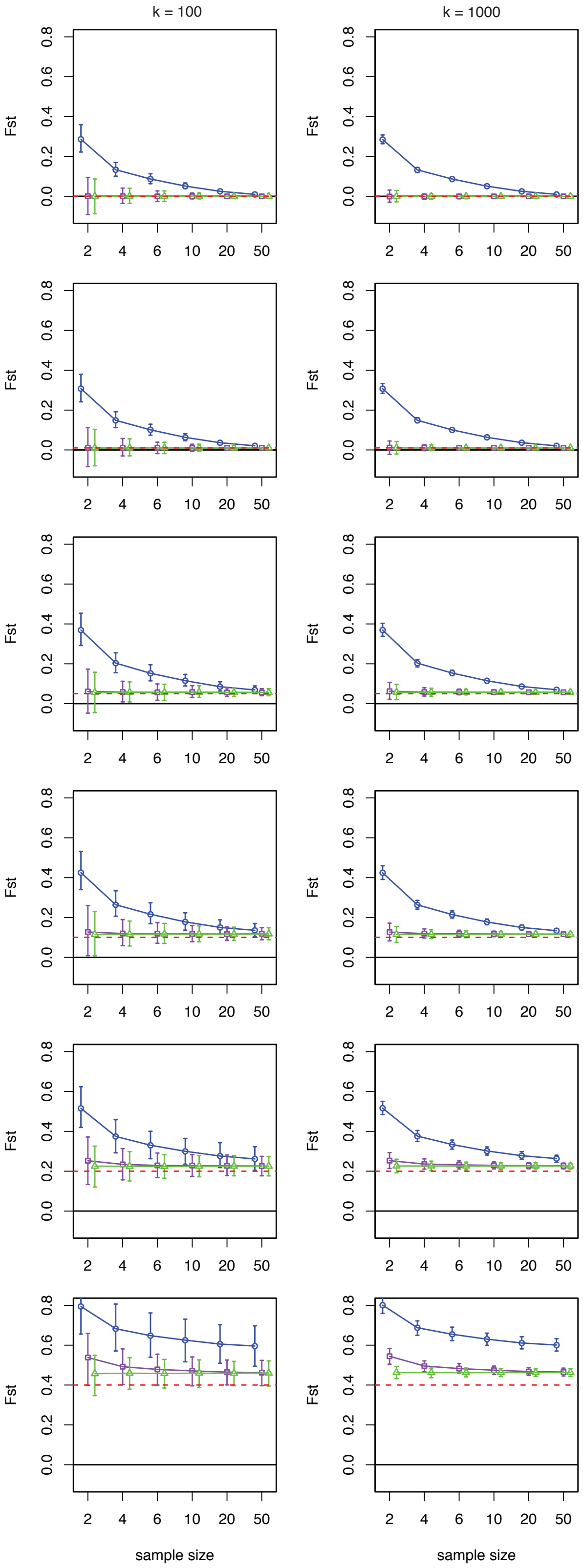


Figure S3.3: Effect of increasing sample sizes. Results are shown for the simulated data with skewed allele frequencies (MAF > 0.25). Number of loci is fixed at k = 100 (left column) and k = 1,000 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: $F_{ST}{}^{W}$ (blue circles), $F_{ST}{}^{W\&C}$ (purple squares) and $F_{ST}{}^{R}$ (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

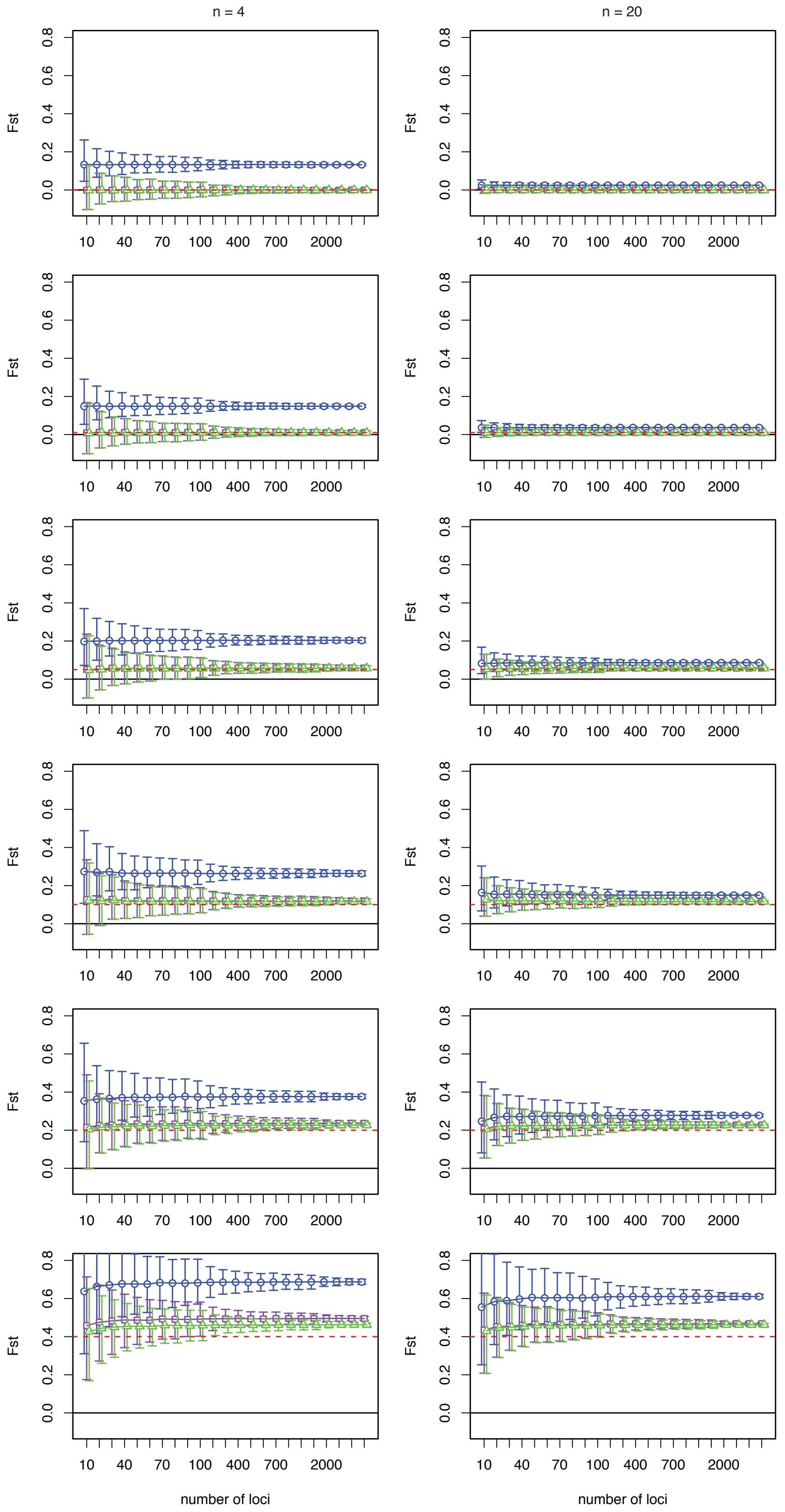


Figure S3.4: Effect of increasing the number of markers. Results are shown for the simulated data with skewed allele frequencies (MAF > 0.25). Number of individuals is fixed at n = 4 (left column) and n = 20 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: $F_{ST}{}^{W}$ (blue circles), $F_{ST}{}^{W\&C}$ (purple squares) and $F_{ST}{}^{R}$ (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

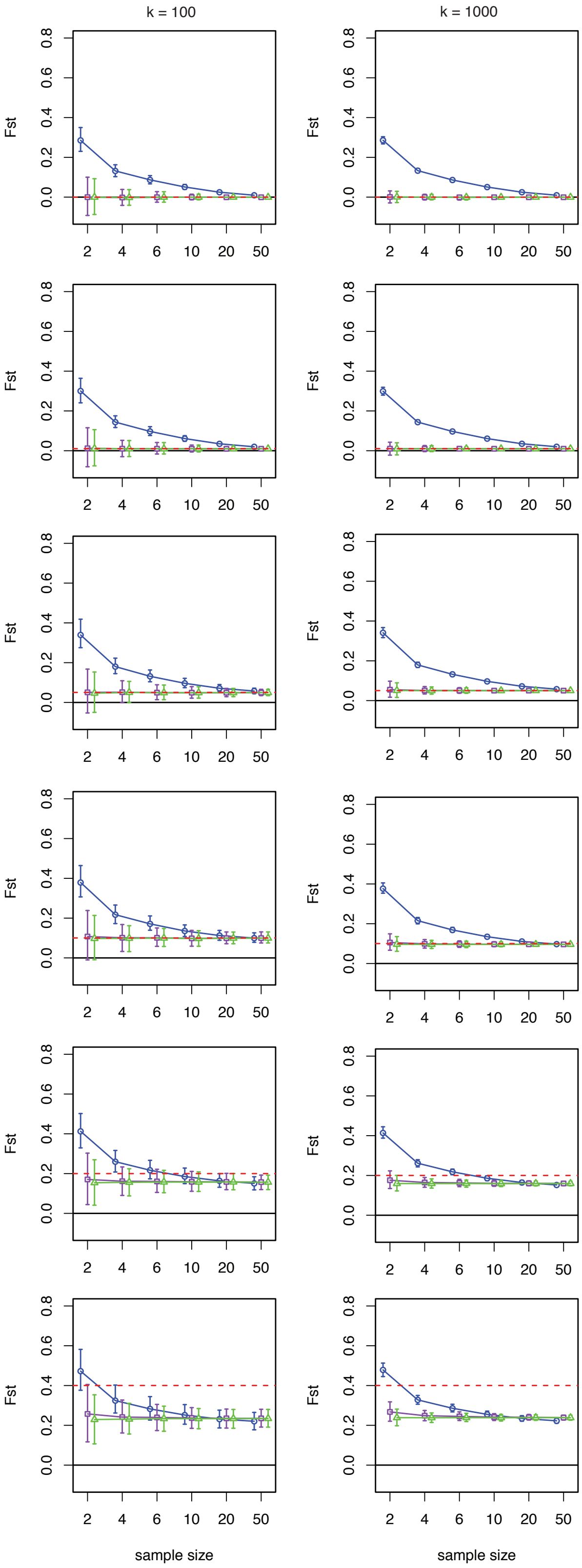


Figure S3.5: Effect of increasing sample sizes. Results are shown for the simulated data with skewed allele frequencies (MAF $\leq$ 0.25). Number of loci is fixed at k = 100 (left column) and k = 1,000 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: $F_{ST}{}^{W}$ (blue circles), $F_{ST}{}^{W\&C}$ (purple squares) and $F_{ST}{}^{R}$ (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

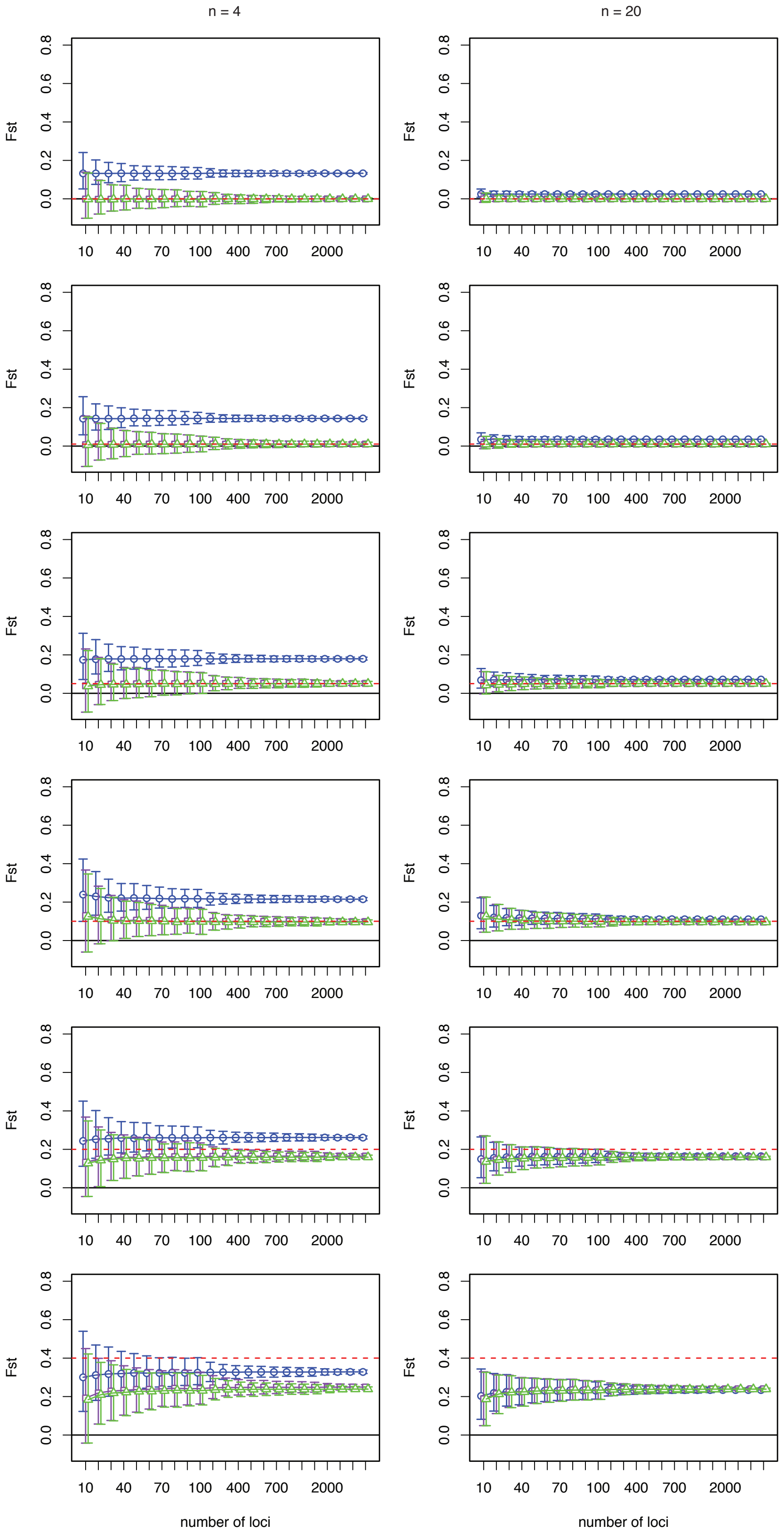


Figure S3.6: Effect of increasing the number of markers. Results are shown for the simulated data with skewed allele frequencies (MAF $\leq$ 0.25). Number of individuals is fixed at n = 4 (left column) and n = 20 (right column). Each row contains a different level of genetic differentiation (F_{ST} = 0, 0.01, 0.05, 0.1, 0.2, 0.4). The results (average F_{ST} and 95% CI) of each estimator are depicted in the different graphs: ${F_{ST}}^{W}$ (blue circles), ${F_{ST}}^{W\&C}$ (purple squares) and ${F_{ST}}^{R}$ (green triangles). The dashed red line indicates the actual F_{ST} for the simulated population.

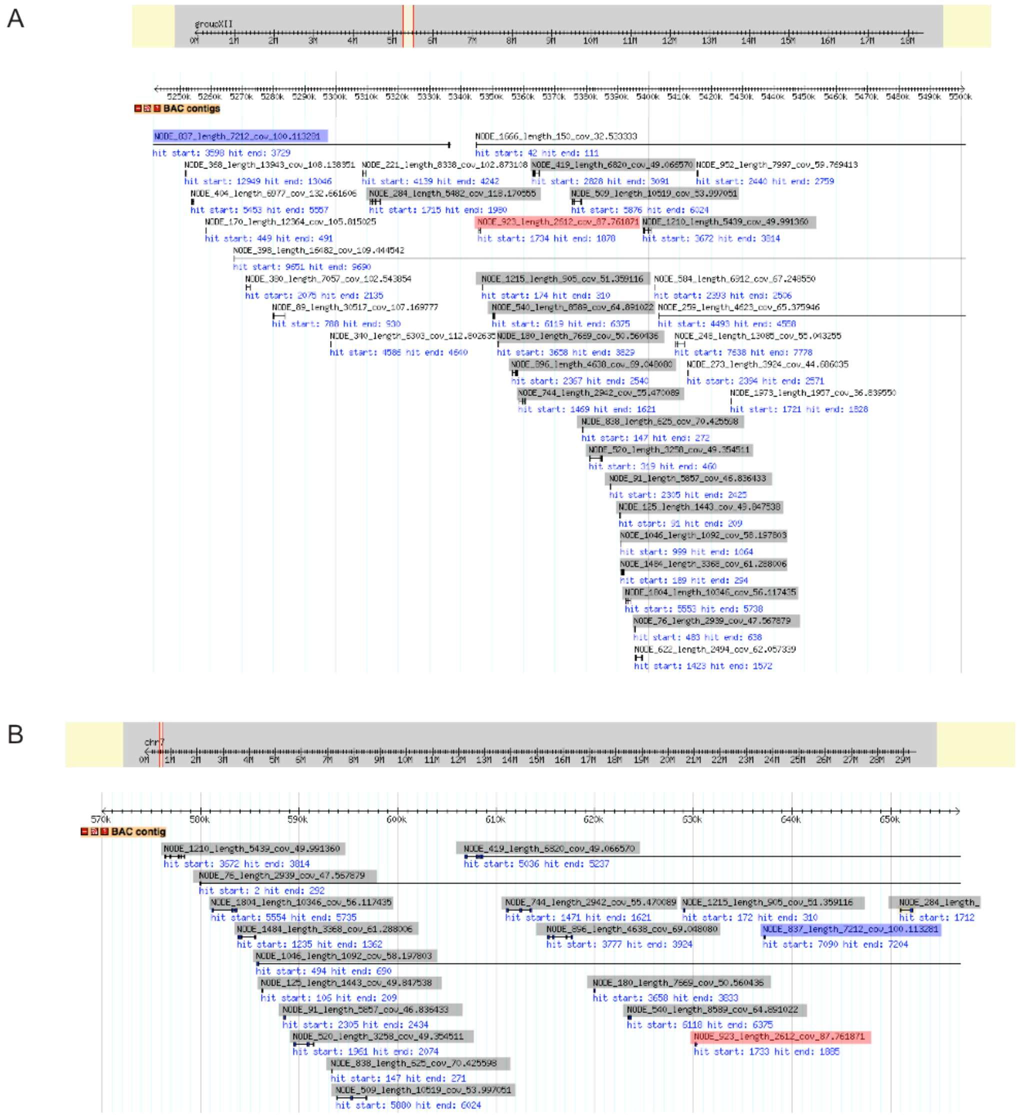

A
groupXII
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M
5250k 5260k 5270k 5280k 5290k 5300k 5310k 5320k 5330k 5340k 5350k 5360k 5370k 5380k 5390k 5400k 5410k 5420k 5430k 5440k 5450k 5460k 5470k 5480k 5490k 5500k
BAC contigs
NODE_837_length_7212_cov_100.113281
hit start: 3598 hit end: 3729
NODE_1666_length_150_cov_32.533333
hit start: 42 hit end: 111
NODE_368_length_13943_cov_108.138351
hit start: 12949 hit end: 13046
NODE_221_length_8338_cov_102.873108
hit start: 4139 hit end: 4242
NODE_419_length_6820_cov_49.066570
hit start: 2828 hit end: 3091
NODE_952_length_7997_cov_59.769413
hit start: 2440 hit end: 2759
NODE_404_length_6977_cov_132.661606
hit start: 5453 hit end: 5557
NODE_284_length_5482_cov_118.170555
hit start: 1715 hit end: 1980
NODE_509_length_10519_cov_53.997051
hit start: 5876 hit end: 6024
NODE_170_length_12364_cov_105.815025
hit start: 449 hit end: 491
NODE_923_length_2612_cov_87.761871
hit start: 1734 hit end: 1878
NODE_1210_length_5439_cov_49.991360
hit start: 3672 hit end: 3814
NODE_398_length_16482_cov_109.444542
hit start: 9651 hit end: 9690
NODE_380_length_7057_cov_102.543854
hit start: 2075 hit end: 2135
NODE_1215_length_905_cov_51.359116
hit start: 174 hit end: 310
NODE_584_length_6912_cov_67.248550
hit start: 2393 hit end: 2506
NODE_89_length_30517_cov_107.169777
hit start: 788 hit end: 930
NODE_540_length_8589_cov_64.891022
hit start: 6119 hit end: 6375
NODE_259_length_4623_cov_65.375946
hit start: 4493 hit end: 4558
NODE_340_length_6303_cov_112.802635
hit start: 4586 hit end: 4640
NODE_180_length_7669_cov_50.560436
hit start: 3658 hit end: 3829
NODE_248_length_13085_cov_55.043255
hit start: 7638 hit end: 7778
NODE_896_length_4638_cov_69.048080
hit start: 2367 hit end: 2540
NODE_273_length_3924_cov_44.686035
hit start: 2394 hit end: 2571
NODE_744_length_2942_cov_55.470089
hit start: 1469 hit end: 1621
NODE_1973_length_1957_cov_36.839550
hit start: 1721 hit end: 1828
NODE_838_length_625_cov_70.425598
hit start: 147 hit end: 272
NODE_520_length_3258_cov_49.354511
hit start: 319 hit end: 460
NODE_91_length_5857_cov_46.836433
hit start: 2305 hit end: 2425
NODE_125_length_1443_cov_49.847538
hit start: 91 hit end: 209
NODE_1046_length_1092_cov_58.197803
hit start: 999 hit end: 1064
NODE_1484_length_3368_cov_61.288006
hit start: 189 hit end: 294
NODE_1804_length_10346_cov_56.117435
hit start: 5553 hit end: 5738
NODE_76_length_2939_cov_47.567879
hit start: 483 hit end: 638
NODE_622_length_2494_cov_62.057339
hit start: 1423 hit end: 1572
B
chr7
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M 24M 25M 26M 27M 28M 29M
570k 580k 590k 600k 610k 620k 630k 640k 650k
BAC contig
NODE_1210_length_5439_cov_49.991360
hit start: 3672 hit end: 3814
NODE_419_length_6820_cov_49.066570
hit start: 5036 hit end: 5237
NODE_76_length_2939_cov_47.567879
hit start: 2 hit end: 292
NODE_1804_length_10346_cov_56.117435
hit start: 5554 hit end: 5735
NODE_744_length_2942_cov_55.470089
hit start: 1471 hit end: 1621
NODE_1215_length_905_cov_51.359116
hit start: 172 hit end: 310
NODE_284_length_
hit start: 1712
NODE_1484_length_3368_cov_61.288006
hit start: 1235 hit end: 1362
NODE_896_length_4638_cov_69.048080
hit start: 3777 hit end: 3924
NODE_837_length_7212_cov_100.113281
hit start: 7090 hit end: 7204
NODE_1046_length_1092_cov_58.197803
hit start: 494 hit end: 690
NODE_125_length_1443_cov_49.847538
hit start: 106 hit end: 209
NODE_180_length_7669_cov_50.560436
hit start: 3658 hit end: 3833
NODE_91_length_5857_cov_46.836433
hit start: 2305 hit end: 2434
NODE_540_length_8589_cov_64.891022
hit start: 6118 hit end: 6375
NODE_520_length_3258_cov_49.354511
hit start: 1961 hit end: 2074
NODE_923_length_2612_cov_87.761871
hit start: 1733 hit end: 1885
NODE_838_length_625_cov_70.425598
hit start: 147 hit end: 271
NODE_509_length_10519_cov_53.997051
hit start: 5880 hit end: 6024

Figure S5.1: Assigning contigs to BACs using GBrowse. All contigs were blasted against the Stickleback and medaka reference genome and alignments were visualized in GBrowse. The search function was used in order find the contigs containing Sanger sequenced parts of the BACs (see Table Sxx[soll das Tab S4.1 sein oder was neues?]). Contigs aligning densely to the same genomic region were believed to belong to the same BAC. This picture shows an example for Marker 0455/0581. Contigs with significant hits on Sanger sequences associated to Marker 0455 are marked in red. Contigs with significant hits on Sanger sequences associated to Marker 0581 are marked in blue. Contigs that have the same ordering on the reference genomes are marked in grey. A) Stickleback B) Medaka

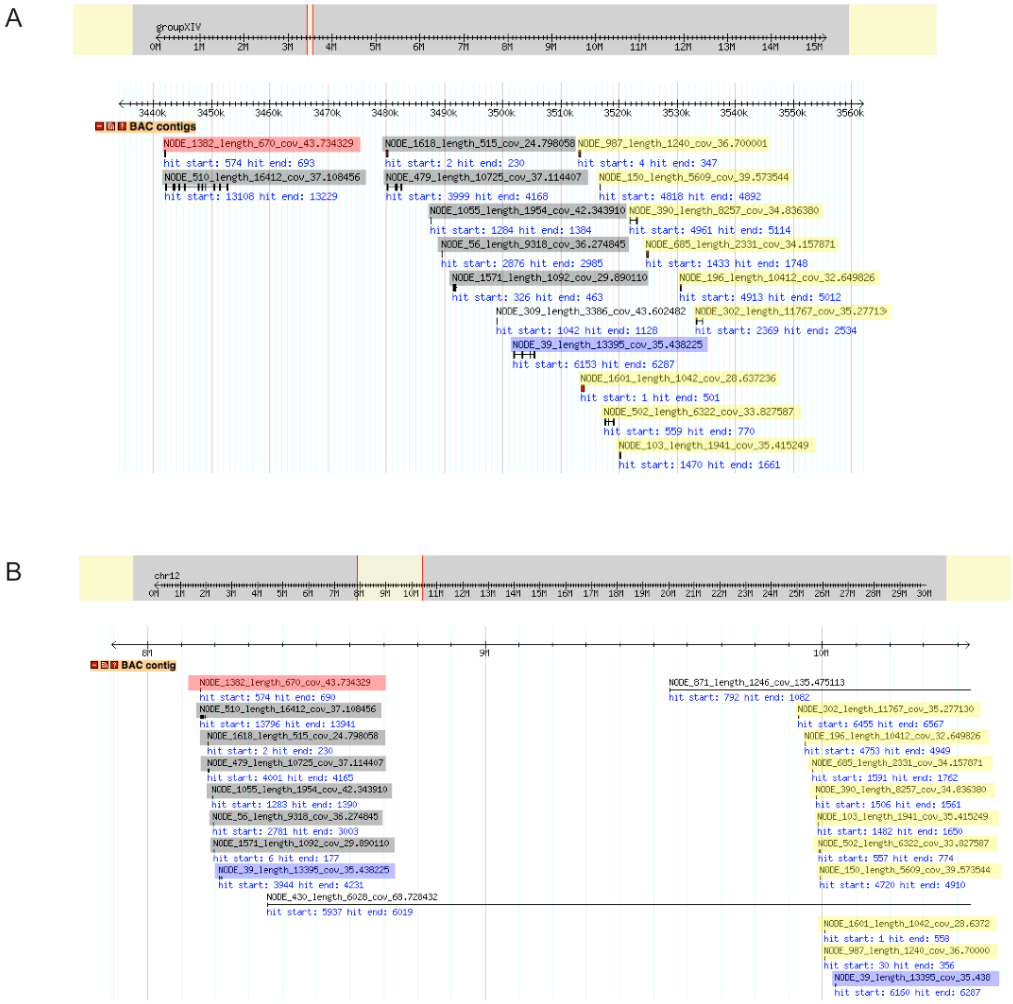

A
groupXIV
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M
3440k 3450k 3460k 3470k 3480k 3490k 3500k 3510k 3520k 3530k 3540k 3550k 3560k
BAC contigs
NODE_1382_length_670_cov_43.734329
hit start: 574 hit end: 693
NODE_510_length_16412_cov_37.108456
hit start: 13108 hit end: 13229
NODE_1618_length_515_cov_24.798058
hit start: 2 hit end: 230
NODE_987_length_1240_cov_36.700001
hit start: 4 hit end: 347
NODE_479_length_10725_cov_37.114407
hit start: 3999 hit end: 4168
NODE_150_length_5609_cov_39.573544
hit start: 4818 hit end: 4892
NODE_1055_length_1954_cov_42.343910
hit start: 1284 hit end: 1384
NODE_390_length_8257_cov_34.836380
hit start: 4961 hit end: 5114
NODE_56_length_9318_cov_36.274845
hit start: 2876 hit end: 2985
NODE_685_length_2331_cov_34.157871
hit start: 1433 hit end: 1748
NODE_1571_length_1092_cov_29.890110
hit start: 326 hit end: 463
NODE_196_length_10412_cov_32.649826
hit start: 4913 hit end: 5012
NODE_309_length_3386_cov_43.602482
hit start: 1042 hit end: 1128
NODE_302_length_11767_cov_35.27713
hit start: 2369 hit end: 2534
NODE_39_length_13395_cov_35.438225
hit start: 6153 hit end: 6287
NODE_1601_length_1042_cov_28.637236
hit start: 1 hit end: 501
NODE_502_length_6322_cov_33.827587
hit start: 559 hit end: 770
NODE_103_length_1941_cov_35.415249
hit start: 1470 hit end: 1661
B
chr12
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M 24M 25M 26M 27M 28M 29M 30M
8M 9M 10M
BAC contig
NODE_1382_length_670_cov_43.734329
hit start: 574 hit end: 690
NODE_510_length_16412_cov_37.108456
hit start: 13796 hit end: 13941
NODE_1618_length_515_cov_24.798058
hit start: 2 hit end: 230
NODE_479_length_10725_cov_37.114407
hit start: 4001 hit end: 4165
NODE_1055_length_1954_cov_42.343910
hit start: 1283 hit end: 1390
NODE_56_length_9318_cov_36.274845
hit start: 2781 hit end: 3003
NODE_1571_length_1092_cov_29.890110
hit start: 6 hit end: 177
NODE_39_length_13395_cov_35.438225
hit start: 3944 hit end: 4231
NODE_430_length_6028_cov_68.728432
hit start: 5937 hit end: 6019
NODE_871_length_1246_cov_135.475113
hit start: 792 hit end: 1082
NODE_302_length_11767_cov_35.277130
hit start: 6455 hit end: 6567
NODE_196_length_10412_cov_32.649826
hit start: 4753 hit end: 4949
NODE_685_length_2331_cov_34.157871
hit start: 1591 hit end: 1762
NODE_390_length_8257_cov_34.836380
hit start: 1506 hit end: 1561
NODE_103_length_1941_cov_35.415249
hit start: 1482 hit end: 1650
NODE_502_length_6322_cov_33.827587
hit start: 557 hit end: 774
NODE_150_length_5609_cov_39.573544
hit start: 4720 hit end: 4910
NODE_1601_length_1042_cov_28.6372
hit start: 1 hit end: 558
NODE_987_length_1240_cov_36.70000
hit start: 30 hit end: 356
NODE_39_length_13395_cov_35.438
hit start: 6160 hit end: 6287

Figure S5.2: Assigning contigs to BACs using GBrowse. All contigs were blasted against the Stickleback and Medaka reference genome and alignments were visualized in GBrowse. The search function was used in order find the contigs containing Sanger sequenced parts of the BACs (see Table Sxx)??. Contigs aligning densely to the same genomic region were believed to belong to the same BAC. This picture shows an example for Marker 1075. Here, all contigs align nearby each other in Stickleback, but are separated if aligned to Medaka. Contigs with significant hits on Sanger sequences associated to Marker 0455 [was tut der hier?]are marked in red. The contig most likely containing the break point is marked in blue. Contigs that have the same ordering on the reference genomes are marked in grey or yellow. A) Stickleback B) Medaka

B. Supplementary Tables

Table S2.1: Pairwise F_{ST}-values calculated A) between the populations within the Northern drainages B) between the populations from the Northern drainages and the populations from the Oropouche drainage C) between the populations from the Northern drainages and the populations from the Caroni drainage D) between the populations from the Northern drainages and the populations from South West Trinidad E) between the populations from the Northern drainages and the populations from Venezuela.

Region	River	Name	Number	1	2	3	4	5	6	7	8	9	10	11
A) Northern drainages	Marianne	M1	1											
		M11	2	0.773										
		M15	3	0.421	0.621									
		M2	4	0.544	0.749	0.188								
		M8	5	0.555	0.633	0.192	0.399							
		MM1	6	0.381	0.552	0.063*	0.168	0.209						
		MM2	7	0.343	0.597	-	0.145	0.216	-					
	Paria	P1	8	0.736	0.541	0.560	0.690	0.609	0.486	0.506				
		P7	9	0.789	0.743	0.661	0.776	0.715	0.595	0.586	0.583			
	Yarra	LY	10	0.598	0.654	0.532	0.582	0.569	0.510	0.499	0.646	0.671		
		UY	11	0.709	0.845	0.672	0.752	0.720	0.635	0.598	0.806	0.833	0.559	
B) Oropouche drainage	Matura	LM	12	0.593	0.608	0.516	0.583	0.560	0.507	0.505	0.578	0.606	0.400	0.623
	Oropouche	LO	13	0.563	0.588	0.506	0.559	0.546	0.500	0.484	0.587	0.585	0.445	0.606
		Oro209	14	0.764	0.779	0.705	0.760	0.749	0.682	0.666	0.760	0.786	0.617	0.771
		Oro201	14	0.704	0.729	0.651	0.701	0.699	0.646	0.637	0.719	0.733	0.559	0.721
		Oro4-2	14	0.615	0.649	0.570	0.625	0.620	0.545	0.526	0.641	0.631	0.479	0.691
		Oro2	14	0.691	0.715	0.638	0.686	0.684	0.621	0.606	0.699	0.712	0.561	0.693
		UO	15	0.586	0.606	0.543	0.591	0.577	0.515	0.492	0.594	0.611	0.457	0.620
	Quare	LQ	16	0.542	0.605	0.504	0.557	0.545	0.484	0.477	0.576	0.587	0.409	0.581
		QuaII6	17	0.781	0.830	0.751	0.792	0.784	0.741	0.731	0.812	0.823	0.705	0.804
		QuaII203	17	0.809	0.844	0.768	0.808	0.799	0.759	0.746	0.828	0.840	0.711	0.827
		QuaII206	17	0.790	0.833	0.756	0.799	0.789	0.754	0.739	0.813	0.826	0.702	0.807
	Turure	LT	18	0.704	0.718	0.627	0.681	0.670	0.608	0.605	0.701	0.726	0.426	0.732
		UT	19	0.747	0.780	0.692	0.745	0.720	0.665	0.662	0.764	0.790	0.483	0.773
C) Caroni drainage	Aripo	LA	20	0.659	0.697	0.608	0.654	0.660	0.582	0.588	0.684	0.707	0.399	0.704
		RP	21	0.700	0.733	0.621	0.691	0.680	0.590	0.596	0.699	0.735	0.418	0.721
		UA	22	0.822	0.869	0.768	0.835	0.818	0.736	0.727	0.837	0.867	0.621	0.859
	Guanapo	LEIC	23	0.798	0.850	0.763	0.812	0.803	0.740	0.743	0.828	0.867	0.562	0.839
		UEIC	24	0.839	0.885	0.785	0.850	0.830	0.766	0.769	0.869	0.895	0.615	0.875
		LG	25	0.679	0.726	0.642	0.688	0.681	0.610	0.610	0.716	0.756	0.415	0.720
		UG	26	0.811	0.862	0.766	0.817	0.807	0.735	0.729	0.852	0.871	0.614	0.870
	Caura	LC	27	0.580	0.619	0.521	0.571	0.558	0.482	0.496	0.601	0.632	0.243	0.593
		UC	28	0.694	0.741	0.607	0.666	0.660	0.583	0.579	0.709	0.737	0.254	0.724
	Tunapuna*	Tu	29	0.660	0.712	0.603	0.663	0.653	0.542	0.573	0.698	0.697	0.377	0.696
	Tranquille*	Tra	30	0.623	0.680	0.574	0.624	0.611	0.553	0.552	0.668	0.686	0.305	0.647
	Lopinot	LL	31	0.578	0.626	0.521	0.572	0.559	0.461	0.468	0.600	0.620	0.294	0.621
		UL	32	0.845	0.895	0.793	0.842	0.836	0.769	0.758	0.876	0.908	0.629	0.880
D) SW Trinidad	Palm Drive	PD	33	0.535	0.579	0.473	0.541	0.512	0.449	0.449	0.562	0.596	0.300	0.585
	Pitch Lake	PL	34	0.775	0.826	0.731	0.777	0.778	0.699	0.709	0.788	0.818	0.605	0.804
E) Venezuela	Cumaná	AP	35	0.805	0.825	0.779	0.807	0.800	0.761	0.763	0.810	0.826	0.745	0.821
		CCB	36	0.723	0.744	0.692	0.728	0.717	0.677	0.671	0.732	0.757	0.632	0.726
		CCE	36	0.764	0.792	0.731	0.766	0.758	0.724	0.714	0.776	0.799	0.691	0.766
		CCFR	36	0.822	0.846	0.807	0.830	0.821	0.793	0.787	0.832	0.854	0.767	0.831
	Poza Azufre	PV6	37	0.732	0.794	0.675	0.728	0.715	0.665	0.650	0.756	0.786	0.555	0.756

NOTE: Populations were defined by sampling site. For all entries $P < 0.001$ if not otherwise marked. $^{**}P < 0.01$, $^{*}P < 0.05$. Not significant estimates are not reported.

Table S2.1 (continued): Pairwise F_{ST}-values calculated A) between the populations within the Oropouche drainage B) between the populations from the Oropouche drainage and the populations from the Caroni drainage C) between the populations from the Oropouche drainage and the populations from South West Trinidad C) between the populations from the Oropouche drainage and the populations from Venezuela.

Region	River	Name	Number	12	13	14	14	14	14	15	16	17	17	17	18	19
A) Oropouche drainage	Matura	LM	12													
	Oropouche	LO	13	0.195												
		Oro209	14	0.408	0.239											
		Oro201	14	0.324	0.135	0.383										
		Oro4-2	14	0.259	0.146	0.341	0.294									
		Oro2	14	0.305	0.133	0.393	0.250	0.261								
		UO	15	0.209	-	0.256	0.196	0.168	0.163							
	Quare	LQ	16	0.232	0.069	0.276	0.206	0.205	0.205	0.084						
		QuaII6	17	0.542	0.367	0.596	0.500	0.497	0.484	0.426	0.332					
		QuaII203	17	0.560	0.407	0.625	0.537	0.525	0.518	0.462	0.349	-				
		QuaII206	17	0.551	0.397	0.617	0.520	0.513	0.482	0.442	0.340	-	-			
	Turure	LT	18	0.432	0.403	0.630	0.559	0.512	0.558	0.441	0.412	0.704	0.718	0.720		
		UT	19	0.506	0.500	0.693	0.633	0.586	0.642	0.504	0.494	0.757	0.775	0.762	0.036	
B) Caroni drainage	Aripo	LA	20	0.441	0.416	0.601	0.564	0.465	0.567	0.454	0.400	0.674	0.700	0.681	0.297	0.363
		RP	21	0.470	0.472	0.650	0.606	0.516	0.594	0.510	0.469	0.715	0.735	0.720	0.353	0.400
		UA	22	0.592	0.571	0.768	0.723	0.637	0.706	0.629	0.575	0.809	0.821	0.812	0.522	0.582
	Guanapo	LEIC	23	0.581	0.555	0.749	0.704	0.639	0.698	0.583	0.550	0.808	0.821	0.819	0.174	0.254
		UEIC	24	0.636	0.587	0.793	0.734	0.673	0.743	0.620	0.582	0.834	0.850	0.844	0.263	0.372
		LG	25	0.464	0.428	0.644	0.599	0.515	0.596	0.479	0.443	0.722	0.735	0.729	0.052**	0.061**
		UG	26	0.589	0.553	0.753	0.713	0.631	0.703	0.612	0.563	0.803	0.811	0.805	0.258	0.263
	Caura	LC	27	0.327	0.371	0.556	0.499	0.416	0.489	0.395	0.359	0.651	0.662	0.659	0.242	0.289
		UC	28	0.444	0.466	0.647	0.601	0.526	0.576	0.486	0.463	0.728	0.740	0.724	0.349	0.378
	Tunapuna*	Tu	29	0.407	0.469	0.647	0.611	0.523	0.594	0.504	0.458	0.724	0.733	0.720	0.376	0.398
	Tranquille*	Tra	30	0.379	0.375	0.569	0.505	0.422	0.528	0.410	0.373	0.669	0.679	0.676	0.327	0.384
	Lopinot	LL	31	0.355	0.383	0.562	0.505	0.449	0.517	0.427	0.381	0.661	0.675	0.664	0.224	0.271
		UL	32	0.618	0.588	0.788	0.746	0.671	0.727	0.655	0.592	0.824	0.842	0.837	0.629	0.690
C) SW Trinidad	Palm Drive	PD	33	0.325	0.304	0.523	0.439	0.375	0.445	0.367	0.342	0.629	0.643	0.638	0.248	0.335
	Pitch Lake	PL	34	0.534	0.547	0.717	0.667	0.604	0.656	0.560	0.539	0.776	0.789	0.773	0.622	0.649
D) Venezuela	Cumaná	AP	35	0.722	0.730	0.801	0.783	0.738	0.774	0.728	0.712	0.836	0.843	0.837	0.761	0.757
		CCB	36	0.609	0.637	0.728	0.709	0.659	0.694	0.635	0.634	0.783	0.795	0.785	0.644	0.644
		CCE	36	0.680	0.683	0.764	0.741	0.692	0.726	0.680	0.677	0.808	0.815	0.810	0.713	0.726
		CCFR	36	0.750	0.753	0.823	0.802	0.765	0.798	0.762	0.742	0.853	0.858	0.854	0.775	0.772
	Poza Azufre	PV6	37	0.551	0.559	0.708	0.678	0.602	0.670	0.587	0.572	0.778	0.791	0.779	0.602	0.640

NOTE: Populations were defined by sampling site. For all entries P < 0.001 if not otherwise marked. **P < 0.01, *P < 0.05. Not significant estimates are not reported.

Table S2.1 (continued): Pairwise F_{ST}-values calculated A) between the populations within the Caroni drainage B) between the populations from the Caroni drainage and the populations from South West Trinidad C) between the populations from the Caroni drainage and the populations from Venezuela.

Region	River	Name	Number	20	21	22	23	24	25	26	27	28	29	30	31	32
A) Caroni drainage	Aripo	LA	20													
		RP	21	0.067												
		UA	22	0.225	0.092*											
	Guanapo	LEIC	23	0.455	0.497	0.695										
		UEIC	24	0.533	0.576	0.788	0.097									
		LG	25	0.285	0.330	0.503	0.123	0.208								
		UG	26	0.483	0.517	0.713	0.465	0.584	0.231							
	Caura	LC	27	0.247	0.253	0.430	0.398	0.447	0.239	0.404						
		UC	28	0.351	0.350	0.548	0.479	0.554	0.315	0.561	0.107					
	Tunapuna	Tu	29	0.361	0.389	0.569	0.490	0.546	0.339	0.533	0.148	0.260				
	Tranquille	Tra	30	0.311	0.318	0.515	0.479	0.535	0.332	0.493	0.122	0.245	0.278			
	Lopinot	LL	31	0.266	0.264	0.442	0.379	0.458	0.232	0.454	0.083	0.154	0.195	0.192		
		UL	32	0.622	0.621	0.834	0.798	0.856	0.632	0.814	0.443	0.592	0.610	0.569	0.347	
B) SW Trinidad	Palm Drive	PD	33	0.299	0.313	0.457	0.411	0.454	0.264	0.421	0.134	0.268	0.296	0.237	0.148	0.511
	Pitch Lake	PL	34	0.563	0.593	0.748	0.737	0.780	0.611	0.776	0.462	0.597	0.564	0.516	0.472	0.816
C) Venezuela	Cumaná	AP	35	0.744	0.754	0.809	0.806	0.818	0.754	0.813	0.699	0.745	0.741	0.729	0.697	0.834
		CCB	36	0.630	0.649	0.731	0.717	0.736	0.648	0.723	0.578	0.628	0.604	0.608	0.558	0.747
		CCE	36	0.689	0.703	0.765	0.762	0.788	0.708	0.769	0.641	0.699	0.689	0.676	0.640	0.793
		CCFR	36	0.774	0.772	0.828	0.816	0.832	0.775	0.826	0.726	0.777	0.763	0.753	0.734	0.845
	Poza Azufre	PV6	37	0.567	0.598	0.724	0.714	0.750	0.607	0.727	0.474	0.572	0.565	0.524	0.501	0.774

NOTE: Populations were defined by sampling site. For all entries $P < 0.001$ if not otherwise marked. $^{**}P < 0.01$, $^{*}P < 0.05$. Not significant estimates are not reported.

Table S2.1 (continued): Pairwise F_{ST}-values calculated between all populations defined by sampling site from South-West Trinidad and Venezuela

Region	River	Name	Number	33	34	35	36	36	36
SW Trinnidad	Palm Drive	PO	33						
	Pitch Lake	PL	34	0.511					
Venezuela	Cumaná	AP	35	0.691	0.771				
		CCB	36	0.563	0.657	0.343			
		CCE	36	0.644	0.734	-	0.298		
		CCFR	36	0.723	0.802	0.260	0.444	0.262	
	Poza Azufre	PV6	37	0.501	0.643	0.716	0.541	0.650	0.734

NOTE: Populations were defined by sampling site. For all entries $P < 0.001$ if not otherwise marked. $^{**}P < 0.01$, $^{*}P < 0.05$. Not significant estimates are not reported.

Table S3.1: F_{ST} estimated with the three different estimators. Expected F_{ST} = 0.001

		F_{ST}^{W}			$F_{ST}^{W\&C}$			F_{ST}^{R}		
k	n	2.5%	97.5%	mean F_{ST}	2.5%	97.5%	mean F_{ST}	2.5%	97.5%	mean F_{ST}
1000	2	0.2918	0.3437	0.3160	-0.0359	0.0447	0.0024	-0.0336	0.0413	0.0021
1000	4	0.1340	0.1606	0.1465	-0.0159	0.0189	0.0007	-0.0157	0.0185	0.0006
1000	6	0.0878	0.1055	0.0959	-0.0091	0.0133	0.0010	-0.0091	0.0131	0.0010
1000	10	0.0519	0.0626	0.0570	-0.0045	0.0077	0.0011	-0.0045	0.0077	0.0011
1000	20	0.0259	0.0317	0.0287	-0.0020	0.0044	0.0011	-0.0020	0.0044	0.0011
1000	50	0.0110	0.0134	0.0121	-0.0001	0.0025	0.0011	-0.0001	0.0025	0.0011
2000	2	0.2966	0.3373	0.3149	-0.0280	0.0375	0.0009	-0.0262	0.0347	0.0008
2000	4	0.1366	0.1595	0.1470	-0.0118	0.0160	0.0013	-0.0116	0.0157	0.0013
2000	6	0.0891	0.1046	0.0961	-0.0068	0.0110	0.0012	-0.0067	0.0109	0.0012
2000	10	0.0529	0.0623	0.0572	-0.0035	0.0070	0.0013	-0.0035	0.0070	0.0013
2000	20	0.0265	0.0313	0.0287	-0.0014	0.0039	0.0011	-0.0014	0.0039	0.0011
2000	50	0.0112	0.0131	0.0121	0.0001	0.0022	0.0011	0.0001	0.0022	0.0011
3000	2	0.2972	0.3355	0.3155	-0.0255	0.0301	0.0016	-0.0238	0.0279	0.0014
3000	4	0.1384	0.1585	0.1470	-0.0094	0.0157	0.0013	-0.0092	0.0154	0.0013
3000	6	0.0896	0.1031	0.0959	-0.0066	0.0096	0.0010	-0.0066	0.0096	0.0010
3000	10	0.0531	0.0617	0.0570	-0.0037	0.0064	0.0011	-0.0036	0.0064	0.0011
3000	20	0.0269	0.0308	0.0288	-0.0009	0.0035	0.0011	-0.0009	0.0035	0.0011
3000	50	0.0112	0.0131	0.0121	0.0002	0.0022	0.0011	0.0002	0.0022	0.0011
4000	2	0.3000	0.3341	0.3158	-0.0232	0.0295	0.0020	-0.0217	0.0274	0.0018
4000	4	0.1385	0.1567	0.1467	-0.0098	0.0150	0.0011	-0.0096	0.0147	0.0010
4000	6	0.0907	0.1031	0.0962	-0.0052	0.0096	0.0013	-0.0052	0.0096	0.0012
4000	10	0.0534	0.0614	0.0571	-0.0032	0.0063	0.0012	-0.0032	0.0063	0.0012
4000	20	0.0269	0.0310	0.0288	-0.0008	0.0037	0.0011	-0.0008	0.0037	0.0011
4000	50	0.0113	0.0130	0.0121	0.0003	0.0021	0.0011	0.0003	0.0021	0.0011
5000	2	0.2995	0.3339	0.3153	-0.0221	0.0276	0.0014	-0.0207	0.0256	0.0013
5000	4	0.1390	0.1572	0.1470	-0.0088	0.0143	0.0013	-0.0086	0.0140	0.0013
5000	6	0.0905	0.1031	0.0961	-0.0057	0.0093	0.0012	-0.0057	0.0092	0.0012
5000	10	0.0537	0.0612	0.0570	-0.0027	0.0056	0.0011	-0.0027	0.0056	0.0011
5000	20	0.0271	0.0309	0.0288	-0.0007	0.0034	0.0012	-0.0007	0.0034	0.0012
5000	50	0.0114	0.0129	0.0121	0.0003	0.0020	0.0011	0.0003	0.0020	0.0011

Table S3.2: F_{ST} estimated with the three different estimators. Expected F_{ST} = 0.01

		F_{ST}^{W}			$F_{ST}^{W\&C}$			F_{ST}^{R}		
k	n	2.5%	97.5%	mean F_{ST}	2.5%	97.5%	mean F_{ST}	2.5%	97.5%	mean F_{ST}
1000	2	0.2817	0.3286	0.3039	-0.0220	0.0475	0.0102	-0.0206	0.0441	0.0095
1000	4	0.1345	0.1591	0.1464	-0.0041	0.0263	0.0101	-0.0040	0.0258	0.0099
1000	6	0.0909	0.1088	0.0991	0.0004	0.0206	0.0104	0.0004	0.0205	0.0103
1000	10	0.0569	0.0684	0.0623	0.0039	0.0169	0.0100	0.0039	0.0168	0.0100
1000	20	0.0328	0.0393	0.0360	0.0069	0.0138	0.0102	0.0069	0.0138	0.0102
1000	50	0.0187	0.0223	0.0204	0.0084	0.0123	0.0102	0.0084	0.0123	0.0102
2000	2	0.2876	0.3239	0.3046	-0.0143	0.0428	0.0116	-0.0134	0.0398	0.0108
2000	4	0.1380	0.1570	0.1469	-0.0012	0.0237	0.0105	-0.0012	0.0233	0.0103
2000	6	0.0928	0.1061	0.0990	0.0029	0.0185	0.0102	0.0028	0.0184	0.0101
2000	10	0.0583	0.0668	0.0624	0.0056	0.0149	0.0101	0.0056	0.0149	0.0101
2000	20	0.0337	0.0386	0.0360	0.0077	0.0130	0.0102	0.0077	0.0130	0.0102
2000	50	0.0192	0.0218	0.0205	0.0088	0.0117	0.0102	0.0088	0.0117	0.0102
3000	2	0.2899	0.3200	0.3041	-0.0106	0.0333	0.0107	-0.0099	0.0310	0.0099
3000	4	0.1393	0.1556	0.1467	0.0014	0.0216	0.0103	0.0014	0.0212	0.0102
3000	6	0.0937	0.1057	0.0991	0.0038	0.0180	0.0104	0.0038	0.0179	0.0103
3000	10	0.0591	0.0663	0.0626	0.0065	0.0147	0.0103	0.0065	0.0146	0.0103
3000	20	0.0340	0.0381	0.0360	0.0080	0.0126	0.0102	0.0080	0.0126	0.0102
3000	50	0.0194	0.0216	0.0204	0.0091	0.0115	0.0102	0.0091	0.0115	0.0102
4000	2	0.2915	0.3184	0.3041	-0.0066	0.0321	0.0109	-0.0062	0.0299	0.0101
4000	4	0.1395	0.1554	0.1468	0.0015	0.0220	0.0105	0.0015	0.0216	0.0103
4000	6	0.0943	0.1051	0.0991	0.0044	0.0175	0.0104	0.0044	0.0173	0.0104
4000	10	0.0594	0.0662	0.0626	0.0068	0.0144	0.0103	0.0068	0.0144	0.0103
4000	20	0.0342	0.0380	0.0360	0.0082	0.0125	0.0102	0.0082	0.0125	0.0102
4000	50	0.0195	0.0215	0.0205	0.0093	0.0114	0.0102	0.0093	0.0114	0.0102
5000	2	0.2926	0.3175	0.3043	-0.0075	0.0327	0.0112	-0.0070	0.0304	0.0104
5000	4	0.1404	0.1552	0.1467	0.0025	0.0210	0.0104	0.0025	0.0206	0.0102
5000	6	0.0947	0.1041	0.0991	0.0052	0.0162	0.0104	0.0052	0.0161	0.0103
5000	10	0.0595	0.0660	0.0625	0.0069	0.0144	0.0103	0.0068	0.0144	0.0103
5000	20	0.0344	0.0378	0.0360	0.0085	0.0122	0.0102	0.0085	0.0122	0.0102
5000	50	0.0196	0.0214	0.0205	0.0093	0.0112	0.0103	0.0093	0.0112	0.0103

Table S5.1: List of Markers and associated BACs. Also listed here are the contigs on which the Sanger sequenced BAC ends had a significant blast hit. In addition it is reported on which chromosome the Sanger sequences had a significant hit on stickleback (*G.a.*) and medaka *(O.l.)* * these Sanger sequences are not BAC ends, but EST, cDNA or genomic sequences that should be found in the corresponding BAC sequence

			Sanger sequence		**Blast hits on contig**				**Blast hits on reference species**	
Marker	**LG**	**BAC**	**name**	**length**	**Name**	**length**	**e-value**	**ID**	***G.a.* LG**	***O.l.* LG**
30	**12**	**BAC15_i02**	yaBac15_I02	955	25	7,444	0	96%	IX (1)	1
			zaBac15_I02	977	29	17,576	0	95%	IX (1)	1
76	**17**	**BAC25_P09**	yaBac25_P09	908	538	3,639	0	96%	III (17)	17
			zaBac25_P09	859	455	1,094	0	98%	-	-
			Blu_Testis_6_B19	1,078	90	7,854	0	94%	III (17)	17
			Marker0076*	604	90		0	97%	III (17)	17
85	**8**	**BAC28_O05**	yaBac28_O05	960	232	1,779	0	100%	-	-
					124	418	0	100%	XII (7)	-
			zaBac28_O05	914	128	4,011	0	98%	XI (8)	8
			Tra_Liver_7-4_G02*	962	236	243	2e-95	96%	XI (8)	8
			pr02_A12*	436	660	331	1e-141	98%	XI (8)	8
280	**20**	**BAC01-2_H02**	yaBac01-2_H02	940	79	8,679	0	98%	XII (7)	20
			zaBac01-2_H02	798	1115	3,053	0	99%	-	
290	**7**	**BAC01_H9**	yaBac01_H09	517	837	7,264	0	99%	XII (7)	7
			zaBac01_H09	109	170	12,416	73-47	98%	XII (7)	-
		BAC32_G02	yaBac32_G02	792	1045	3,840	e-134	91%	-	-
			zaBac32_G02	756	923	2,664	0	100%	XII (7)	7
380	**12**	**BAC03_F10**	yaBac03_F10	917	421	2,327	0	92%	-	1
					962	1,314	0	89%	IX (1)	14
					483	801	0	87%	-	23
					48	411	0	98%	-	-
					1172	516	0	92%	-	15
			zaBac03_F10	882	359	2,056	0	93%	XX (16)	-
					626	1,981	0	87%	-	-
396	**8**	**BAC03_I24**	yaBac03_I24	974	416	4,744	0	97%	XI (8)	8
			zaBac03_I24	759	1560	594	0	95%	-	8
455	**1**	**BAC04_G12**	yaBac04_G12	1,012	4	13,059	0	95%	IX (1)	1
			zaBac04_G12	976	152	2,402	0	96%	IX (1)	1
581	**1**	**BAC33_A19**	yaBac33_A19	714	not found					
			zaBac33_A19	445	491	3,690	3e-95	98%	-	1
1075	**12**	**BAC19_N24**	yaBac19_N24	712	1501	695	0	99%	-	
			zaBac19_N24	516	1382	722	0	88%	XIV (12)	12
			C1qTNF3*	341	302	11,819	0	98%	XIV (12)	12
			Rxfp3*	650	1601	1,094	0	99%	XIV (12)	12
			skiv2l2 fw*	204	1382	722	2e-94	100%	XIV (12)	12
			skiv2l2 rev*	169	510	16,464	1e-74	99%	XIV (12)	12
			ppap2a fw*	920	510	16,464	0	99%	XIV (12)	12
			ppap2a rev*	959	510	16,464	0	96%	XIV (12)	12
			aim1_5end*	1,025	657	526	0	97%	-	-
			aim1*	2,013	685	2,331	0	100%	XIV (12)	12

C. RApiD Manual

This is the manual for the pipeline RApiD (Willing, et al., 2011) to analyze paired-end RAD-seq data. The following external tools are required. The output of these tools is used as input for the scripts that are part of RApiD.

1. The assembler LOCAS available from http://ab.inf.unituebingen.de/software/locas/, in order to assemble the 2nd read clusters. We developed as addition LOCASopt (part of RApiD) in order to optimize each local assembly.
2. The read aligner and clustering algorithm vmatch in order to cluster read1. Available under www.vmatch.de
3. The read aligner GenomeMapper (Schneeberger et al., 2009, http://www.1001genomes.org/downloads/genomemapper_singleref.html), in order to map the reads back to the assembled reference.

Convert Fastq to Fasta
Usage: Fastq2Fasta.pl –i <name>.fastq
-i fastq file that should be converted to fasta

Generates a fasta output file called <name>.fasta

Parse clustering output from Vmatch
Usage: parseVmatchClustering.pl -i input-file -l input-file -o output-file

prints the cluster number, size and read identfiers of each cluster in the output file
1 line per cluster

-i fasta file containing the reads that were used as input for vmatch
-v output file from vmatch
-o cluster file that can be used as input the the assembly script

Vmatch commands:

```
mkvtree –db read1.fasta –dna –pl -allout
vmatch -d -l 84 -dbcluster 100 100 -h 3 -v read1.fasta
```

Assembly of read2 clusters
In order to run the assembly go to http://ab.inf.uni-tuebingen.de/software/locas/ and download LOCAS. Move the binary LOCASopt, which is provided with the package RApiD to the locas directory. You need to give the full path to LOCASopt to the script.

Usage: assembleRead2Clusters.pl -i input-file -v input-file -o string –c int –r int –n integer -a string -k int-int-int –l int-int-int –e float-float-float

Goes through the cluster file and calls LOCASopt on each cluster separately. Assembled contigs are already reverse complemented and therefore have the same strand direction as read1.

-i input file with second illumina reads in fasta format

-v clusters (output from parseVmatchClusters.pl)
-o beginning of output files, 2 files are generated <output>_contigs.fasta and <output>_ids.fasta
-c maximal number of contigs allowed in assembly
-r range of #reads required in clusters, min_number – max_number, e.g. 5-192, all clusters with a size of 5 to 192 reads are assembled
-a path to LOCASopt
-k kmer range tested with LOCASopt, minimal kmer – maximal kmer – size of steps
e.g. 13-17-2, means kmers 13, 15 and 17 are used
-l range of overlap length tested, minimal overlap length – maximal overlap length – size of steps, e.g. 21-27-2, means overlap lengths 21, 23, 25 and 27 are used
-e error rate allowed in overlap, minimal error rate – maximal error rate – size of steps, e.g. 0.05-0.07-0.01 means error rates 0.05, 0.06 and 0.07 are tested

Join read1 and read2 contig
Usage: joinRead1WithRead2contig.pl -i input-file -a input-file -r input-file -o string –f int –e float

-i id file (output from assembleRead2Clusters.pl), contains the ids of the reads used for each contigs
-a read2 contig file (output from assembleRead2Clusters.pl), contains the assemble read2 contigs
-r fasta file containing read1 of each pair
-o output fasta file, containing the read1 consensus joined with the consistent read2 contig
-f (optional, default = 0, meaning no overlap) if the script should check for an overlap between read1 consensus and read2 contig, set the minimal overlap length
-e (optional, default = 0.0) maximal rate of mismatches allowed in the overlap. Has to be between 0 (0%) and 1 (100%).

Joins each read2 contig with the corresponding read1 consensus and checks for overlap if wanted. If both parts do not overlap, a separator of 10 Ns is inserted between the two parts.
Three output files are generated:
<output>.fasta contains all joined sequences
<output>_notOverlapping.fasta contains all read2 contigs that did not overlap with the corresponding read1 consensus
<output>_Overlapping.fasta contains all overlapping contigs

Sort barcoded reads
Usage: ./countAndSort_BarcodesAndRestrictionSites.pl -1 input-file-fastq [-2 input-file-fastq] -o output-file-txt -f input-file-txt -r input-file-txt –e int [-c 1 -s int]

-1 fastq file with read1
-2 (optional) fastq file with read2, if paired end sequencing
-o output file containing the barcode and restriction site counts

-f file containing the names of the barcodes and the sequence information separated by a tab, e.g.

index1 ACTG
index2 GTCA

-r file with sequence information of restriction site overhang, e.g. for EcoR1 it would be AATTC same format as for barcodes

EcoR1 AATTC
NdeI GTCCT

-e maximal number of errors allowed in barcode

-c (optional, default = 0), if set to 1, barcode and restriction site are removed from read1

-s remove all read1 containing more than a given number of Ns (uncalled nucleotides)

This script counts the occurrences of the different barcodes and sorts the reads into different fasta files named <code>.fasta.
If wanted, it cuts the barcode and restriction site from read1.

Call Consensus

Usage: ./callConsensus.pl -1 input-file -r input-file -l int [-2 input-file -p input-file -m int] -c input-file -o string [-q int -n int -u -x int]

-1 read1-mapping file (output from GenomeMapper, has to be sorted according to the first column (hit names))

-r fastq file containing read1

-l length that should be used from read1 (it might be good to cut the last few bases)

-2 (optional) read2-mapping-file (output from GenomeMapper)

-p (required if -2 is set) fastq file containing read2

-m (required if -2 is set) length that should be used from read2

-d (required if -2 is set) direction of read2 that were mapped (P = original read2 direction, D = reverse complements of read2 were mapped). If you reverse complement the read2 before mapping, you don't have to use the reverse complement of the reference

-c fastq file containing the reference sequence

-o string that gives the beginning of the output file generated (consensus fastq, alignment file, statistics file)

-q (optional, default = 6) min_quality of bases used to call consensus

-n (optional, default = 1) min_coverage to call a consensus base (default = 1)

-u (optional, default = no iupacs) iupac flag, if set, consensus contains iupa codes, if not, the majority base is called as consensus

-x (optional, default = 64), quality format used 64 = illumina, 33 = sanger

Calls the consensus with reads mapped back to a given reference (reads currently only GenomeMapper format).
Three output files are generated:

<output>.fastq contains the consensus sequences in fastq format. Quality values are the average of the quality values of the bases that mapped to this position.

<output>_alignments.txt contains the alignments generated to call the consensus (can be quite big)

<output>_counts.txt contains the coverage per position over all tags

Call SNPs

Usage: ./callSNPs.pl -1 input-file -r input-file -l int [-2 input-file -p input-file -m int] -c input-file -o string [-f float -q int -n int -k int -x int]

-1 read1-mapping file (output from GenomeMapper)

-r fastq file containing read1

-l length that should be used from read1 (it might be good to cut the last few bases)

-2 (optional) read2-mapping-file (output from GenomeMapper)

-p (required if -2 is set) fastq file containing read2

-m (required if -2 is set) length that should be used from read2

-c fastq file containing the reference sequence

-o string that gives the beginning of the output file generated (consensus fastq, alignment file, statistics file)

-f (optional, default = 0.0) minimal minor allele frequency (MAF) to call a SNP

-q (optional, default = 20) minimal quality of bases used to call a SNP

-n (optional, default = 10) minimal coverage for a site containing a SNP

-k (optional, default = 2) minimal number of reads to cover a polymorphism

-x (optional, default = 64) quality format 64 = illumina, 33 = sanger

Checks for SNPs within reads mapped back to a given reference (reads currently only GenomeMapper format). A SNP has to be covered by at least three unique read pairs. Four output files are generated:

<output>.txt contains the SNP information

<output>_alignments.txt contains the alignments generated to call the SNPs (can be quite big)

<output>_counts.txt contains the total coverage of all positions that were checked for a SNP over all tags

<output>_statistics.txt contains read counts used for SNP calling, total #SNPs found etc.